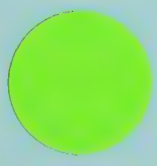
Health Care Financing

Research & Demonstration Series



Report No. 10

A Study of the Classification of Hospitals for Prospective Reimbursement

Department of Health, Education & Welfare • Health Care Financing Administration

A STUDY OF THE CLASSIFICATION OF HOSPITALS

FOR PROSPECTIVE REIMBURSEMENT

Final Report
Submitted to the Office of Research and Statistics
Health Care Financing Administration
U. S. Department of Health, Education, and Welfare

by

T. Klastorin**
C. Watts*
V. Trivedi*

This report is made pursuant to Contract #600-76-0143. The amount charged to the Department of Health, Education, and Welfare for the work resulting in this report (inclusive of the amounts so charged for any prior reports submitted under this contract) is $168,000. The names of the persons, employed or retained by the Contractor, with managerial or professional responsibility for such work, or for the content of the report are listed above.

*School of Public Health and Community Medicine
**Graduate School of Business Administration
University of Washington
Seattle, Washington 98195

January, 1978

TABLE OF CONTENTS

PREFACE

This report represents the findings of a fifteen month research study conducted at the University of Washington under the auspices of the Department of Health Services (School of Public Health and Community Medicine) and the Graduate School of Business. While this study dealt primarily with the problem of developing a viable procedure for classifying the nation's short-term general hospitals for use in a hospital prospective reimbursement system, a significant number of related issues and topics had to be investigated and studied. Among these were such issues as the economics of hospitals and prospective reimbursement, multivariate statistical methodologies for determining and assessing hospital clusters, and computer software techniques and programs. Extensive efforts were made throughout this study to insure that all questions investigated were studied in a most thorough manner possible; our analysis was based on state-of-the-art methods, and our results and findings were based on the strongest possible foundation.

With any research study of this type, however, advances are made rapidly and new hypotheses are created and old hypotheses are discarded as empirical results dictate. In addition, it must be recognized that this study was constrained to an investigation of the immediate questions posed by the funding agency, and limited by the nature of the data base. In the continuation study proceeding presently, many of these constraints have been relaxed. Therefore, this study should be viewed as a forerunner to the ongoing studies and this report should be considered as the first of several volumes.

Project Director for this study was Professor W. L. Dowling. Co-investigators were Professor T. D. Klastorin (who directed the development of the statistical methodology and empirical investigations) and Professor V. Trivedi. The Project Associate was Professor C. A. Watts, who directed the development of the general economic framework described in Chapter Two. Mr. R. Lanier, the Project Research Assistant, conducted much of the empirical analysis with the assistance of Mr. R. Flewelling, who wrote most of the computer programs for the classification analysis.

The SPSS (Statistical Package for the Social Sciences) program was used for most of the standard statistical analyses, including the factor, regression, and discriminant analyses. This package was initially developed at Stanford University and later developed by the National Opinion Research Center at the University of Chicago. The current version used in this study was developed by the Vogelback Computing Center, Northwestern University, and supported by the University of Washington Computer Center.

The authors gratefully acknowledge the helpful comments, suggestions, and criticisms received from Mr. J. Pettengill during the course of this study, and the secretarial assistance of Ms. C. Sakai and Ms. J. Davis. This study was supported by Contract #600-76-0143 from the Office of Policy, Planning, and Research, Health Care Financing Administration, U. S. Department of Health, Education, and Welfare.

FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION AND OVERVIEW

I. Introduction

The rise in insurance coverage in recent years has been accompanied by a tremendous, and perhaps not unrelated, rise in hospital expenditures. In the resulting search for ways to reduce costs, third parties (including the federal government) have become interested in the concept of prospectively-determined payment rates as a replacement for reimbursement on the basis of reported costs.[1] The major purpose of this study is to examine a method of hospital classification which might be used as part of a prospective reimbursement system.

The impetus for a prospective system comes from the belief that some fraction of the observed cost increase is "unjustifiable." That is, rather than moving from point A in Figure 1.1 to point B as demand responds to rising income and increased insurance coverage, proponents of incentive reimbursement argue that the industry has instead moved from point A to point C. The upward shift in the long run average cost curve (LRAC) is attributable, according to this argument, to two main sources: decreases in the operating (technical) efficiency of hospitals, and changes in the nature or quality of the output (product choice efficiency).[2,3]

It is argued that the first, operating inefficiency, has been encouraged by the payment mechanism by which providers have typically been reimbursed for services rendered. Providers determine service cost retrospectively and are reimbursed for the full amount. As third party coverage increases, the argument goes, this open-ended contract reduces the incentive for hospitals to operate at minimum cost, and the long run average curve moves upward.

---

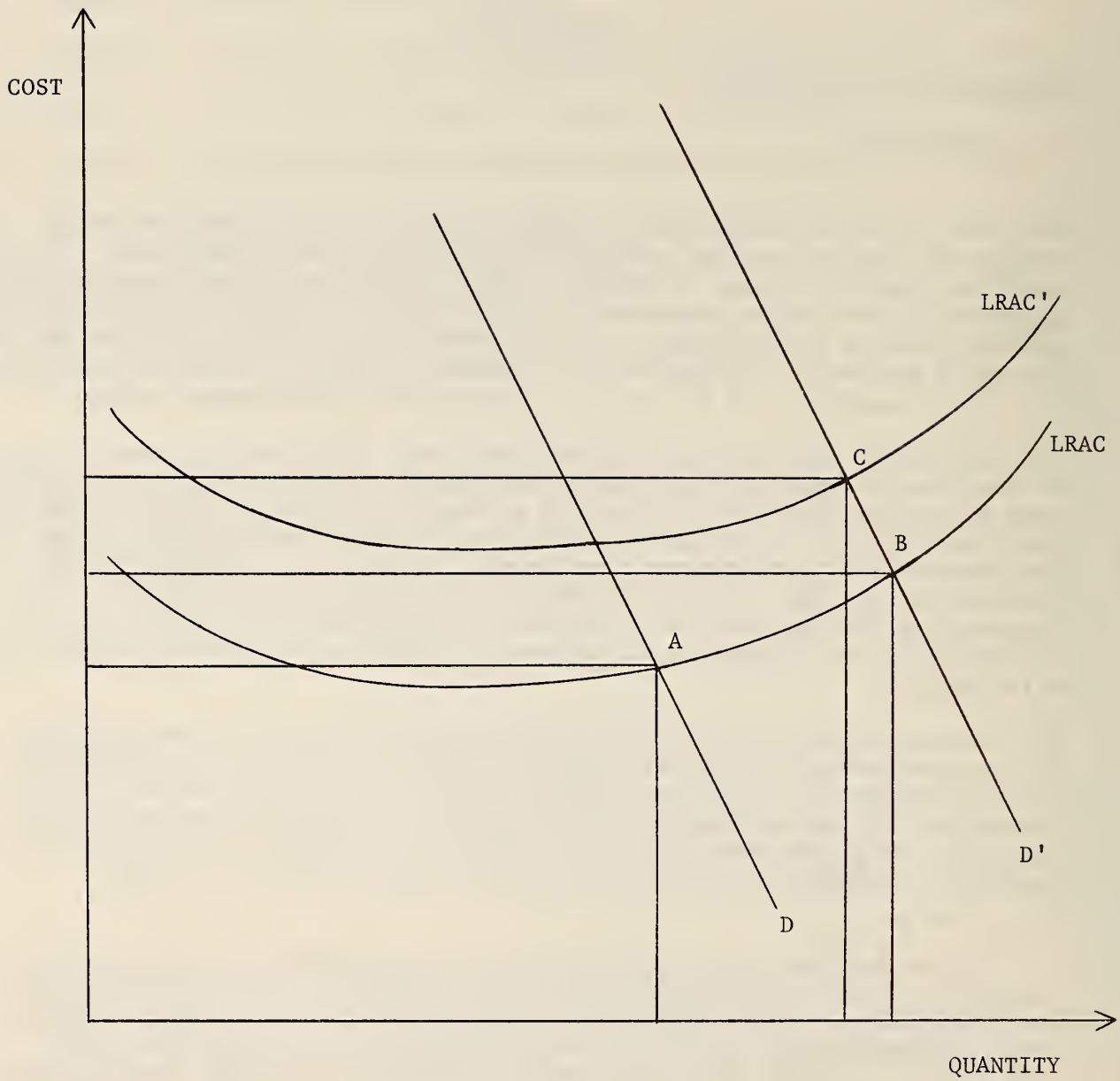
[1] It is interesting to note that the original Medicare law called for reasonable full-cost reimbursement with an additional plus factor for growth and development. Almost immediately, the plus factor was dropped; somewhat later the wording was amended to read reimbursement of "reasonable" cost. Some forms of prospective reimbursement (the name is a slight contradiction of terms) can be seen as an attempt merely to agree upon what constitutes "reasonable cost" before the fact.

[2] Increases in input prices would also cause the curve to shift upward.

[3] Quantity in Figure 1.1 is defined in terms of a standard unit (e.g., hernia equivalent) to avoid measurement problems.

FIGURE 1.1

Health Care Cost Curves



There is some disagreement as to the causes of the second source, changes in the nature or quality of output. It is possible that such changes are the natural responses to increased standards of living, increased technology (i.e., changes in the range of products that can be demanded), and increased insurance coverage.[4] However, some authors[5] argue that many of the changes are strongly influenced by providers--especially physicians--if not actually physician (or hospital) generated, and thus are not reflections of consumers' "true" preferences. Given the level of consumer knowledge in this industry and the resulting dependence of the consumer on the provider's advice, the potential for physician and/or hospital influence is increased as insurance lowers the marginal cost to the patient of additional services, since at the same time it reduces the benefit (in terms of dollar savings of "unnecessary" procedures not performed) of acquiring additional information.

A system of prospectively determined rates is thought to address these two sources of cost increases. If the prospective rate puts the provider at risk for production costs exceeding the agreed-upon rate, the incentive for the institution that wishes to break even (or earn a surplus) is to hold expenditures to this level (or below if surpluses need not be returned to the payor). To the extent operating inefficiencies exist, their elimination is one means of reducing expenditures.

The effect of incentive reimbursement on changes in the nature of output depends upon the degree to which output is provider (hereafter to include physicians as well as hospitals) influenced. The reimbursement mechanism can affect the output vector by changing the incentives facing providers. This point will be discussed more fully in later chapters.

The success of such reimbursement schemes in achieving their cost containment goals without unduly distorting the industry, however, depends crucially on the answers given to several important questions concerning the design of the reimbursement program. How the program parameter issues are resolved will shape the incentives facing providers, and in turn, providers' responses to the program.

The first decision requires determining what payment unit should be used. That is, should the rate be based on units of output: actual services rendered (e.g., one rate for lab test A, another for x-ray B, another for drug C, etc.), days of care (e.g., all-inclusive per diem rate), or admissions (all-inclusive or diagnostic specific rates); or should it be based on units of time: yearly (monthly) departmental or institutional budgets? The choice of this parameter has a direct impact on the issue of provider influence of demand since it determines to a large extent the financial desirability to the hospital of changing the output vector.

---

[4] Undesirable changes in consumer demand brought about by insurance through the presence of moral hazard are best addressed by changing the nature of the insurance contract. Altering the provider payment mechanism will have little impact on the consumer-generated demand curve.

[5] See, for example, Feldstein (1974).

The second decision requires determining which costs should be included in the determination of the prospective rates. That is, should rates cover all operating expenses, or should they apply only to selected pieces of the hospital's operation? For example, in its payment of medicare claims, the Social Security Administration currently applies a pre-determined ceiling only to routine costs, leaving all other costs (ancillary services, drugs, etc.) payable retrospectively. The choice of this parameter is important if hospitals are to be prevented from escaping effective control by reallocating costs or from distorting the output vector by reallocating effort to non-covered services or components.

The final question concerns determination of the set of hospitals upon which a given rate will be imposed. That is, should each provider face a different rate; should all providers face the same rate; or should different rates be set for distinct subgroups of the provider population? Since prospective rates function as simulated prices, it is important that a proper response be given to this question to avoid giving improper signals to providers.

Failure to give careful consideration to each of these questions (summarized in Table 1.1) casts doubts on the desirability and efficacy of a system of incentive reimbursement. This study focuses on an examination of the third question: the criteria by which hospitals should be placed in subgroups for purposes of prospective rate determination.[6] However, the three questions are interrelated to the extent that a discussion of the third cannot be adequately carried out without references to the first and second. Therefore, while the first two questions are not directly within the scope of this study, some attention will be given to these points as they relate to the grouping issue.

TABLE 1.1:

Elements of a Prospective Reimbursement System

1. Payment Unit
   - Per Service
   - Per Day
   - Per Admission
   - Per Unit Time
2. Included Costs
   - Total Costs
   - Routine Costs
3. Coverage of Rates
   - Each Hospital
   - All Hospitals
   - By Subgroups

---

[6] In addition, the classification of hospitals is a crucial aspect of many performance evaluation systems, as suggested by several proposals for establishing a formal system of National Health Insurance (e.g., the McIntyre Bill S.1100, or the Kennedy-Mills Bill - S.3286 and H.R. 13870).

A. Previous Studies

Although hospital classification has been frequently discussed in relation to various incentive reimbursement schemes, there has been only a limited amount of empirical work done to date. One characteristic study, performed by Berry (1973), used data from the American Hospital Association (AHA) and a sample of 4,814 hospitals. Berry analyzed the frequency distribution of services and facilities within hospitals and found five distinct groups of hospitals based on the range of services provided which extended from the most basic services provided in small rural hospitals to the most complex services provided in large metropolitan medical centers (Berry's labels are: Basic, Quality Enhancing, Complex, Community, and Special). Significant differences were found to exist in length of stay, per diem cost, and occupancy rates among these groups of hospitals.

More recent studies have incorporated a larger spectrum of variables for hospital classification. Phillip and Iyer (1975) classified 5,700 hospitals from the AHA Annual Survey of Hospitals, using a total of seventeen "product characteristic" variables (for example, number of beds, number of RN's and LPN's, etc.). Seventy-one hospital groups were generated from the sample using cluster analysis techniques.

Another study, performed by Trivedi (1977), used four categories of variables including hospital demand and supply measures, measures of composition (mix) of hospital output and measures of quantity of output. The study was performed for 94 short-term general hospitals in Washington State and the five groups generated are currently used by that state in its hospital rate review process.

The primary limitation of these studies is that little attention is given to the selection of the classification variables, which is of utmost importance to the success of any payment system based on the resultant groups. Furthermore, the determination of an appropriate statistical classification methodology is nowhere carefully considered. It is the aim of this study to begin to attempt to overcome these shortcomings.

II. Classification Methodologies

The goal of any classification analysis is to analytically classify hospitals in such a way as to simultaneously maximize hospital homogeneity within groups and hospital heterogeneity between groups (i.e., to group hospitals such that hospitals in the same group are more alike than hospitals in different groups). However, if homogeneous clusters of hospitals are not to be defined by an arbitrary process, the method used must resolve a number of crucial questions in a statistically satisfactory manner. Among these, the most important questions which must be considered by any classification methodology are listed below:

1. How can homogeneity among hospitals be precisely and mathematically defined?

2. How many groups should there be?

3. How should weights on each variable be determined?

4. How should the tradeoff between the number of groups and overall homogeneity be made; i.e., if more or fewer groups are desired, how should groups be combined or split in order to meet the desired criterion?

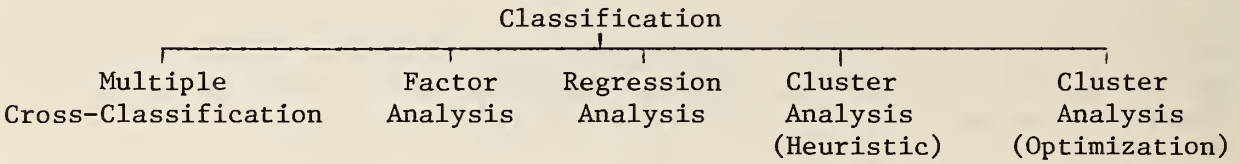
5. How can resultant clusters be validated?

In addition to providing answers to these questions, any viable grouping methodology must meet the usual statistical criteria of efficiency, sufficiency, and consistency.

While a number of methodologies have been proposed for classifying multivariate data (i.e., hospitals described by a vector of characteristics), each has one or more limitations with respect to the questions and/or criteria proposed above. Peterson (1971) points out that many of these methodologies exist on a continuum--at one end are the highly pragmatic techniques of multiple cross-classification, and at the other end there are clustering techniques based on global optimization criterion. As illustrated in Figure 1.2, the techniques of regression analysis, factor analysis, and heuristic cluster analysis lie between the two extremes.

FIGURE 1.2:

Classification Methodologies



The most pragmatic technique of multiple cross-classification is illustrated by the present Health Care Financing Administration (Medicare Program) system of classifying hospitals on the basis of three dimensions:[7] urban-rural designation, state per capita income, and hospital size in bed capacity. The limitations of such a scheme are readily apparent. First, since it is necessary to define variables by interval (as opposed to the precise value), useful information is lost. Further, the designation of the interval range to be used must be arbitrary and cannot be tested against an alternative range. Finally, adding additional measurements will drastically increase the number of associated classification cells and the number of clusters, thus reducing the possible number of hospitals in each cluster. With the exception of the sufficiency criteria, it is apparent that such a scheme fails to satisfactorily resolve the questions stated above.

---

[7] For a more complete discussion of the limitations inherent in the Social Security Administration's present classification system, see Phillip and Iyer, 1974; and Pointer and Phillip, 1974.

Regression analysis has infrequently been proposed as a clustering tool. Having selected relevant independent variables, a regression equation can be determined for a single dependent variable (presumably, in this case some function of hospital cost). Subsequently, the dependent variable values predicted by the resultant equation can be used to determine clusters; since the classification problem has been reduced to a unidimensional problem, properties and procedures determined by Fisher (1958) for finding optimal clusters could be employed for hospital classification. The major limitations of such an approach are threefold: (1) the determination of ultimate hospital groups is based upon the assumptions of linearity and additivity (or whatever model was used in the regression analysis) and might not reflect the true underlying relationship, (2) the variables are assumed to be normally distributed, and (3) all hospital groups are assumed to have the same cost structure. In addition, the dependent variable (i.e., cost) is exceedingly costly to measure and, even if measured, most likely would not reflect efficient cost levels. It would seem more prudent to determine hospital clusters from the independent variables alone, and subsequently retain the capability to develop a reimbursement scheme on whatever statistic or formula seems most appropriate.

The use of factor analysis to classify hospitals suffers from the same weaknesses inherent in regression analysis--the relationship among the variables is constrained to be linear, all hospitals are assumed to face similar cost structures, and the variables are assumed to be normally distributed. The use of this technique, however, for preprocessing variables prior to hospital classification is discussed in Chapter Three.

The last techniques shown in Figure 1.2 are those of heuristic and optimization cluster analysis. These procedures usually combine the vectors of hospital characteristics into a single composite measure between all pairs of hospitals (called a similarity measure) which represents each hospital pair's affinity, and subsequently use this similarity measure to develop some set of reasonable clusters.

In addition to the methodologies indicated in Figure 1.2, there are other multivariate statistical procedures related to the classification problem. For example, discriminant analysis assumes that populations have been identified beforehand and, as most commonly used, finds a linear function which attempts to explain the maximum difference between populations. While discriminant analysis might theoretically be applied to all possible partitions of hospitals, the enormous number of possible groups obviously precludes such effort (this is discussed in further detail in Chapter Three).

Another technique for finding homogeneous clusters is the Automatic Interaction Detector (AID), initially proposed by Morgan and Sonquist (1963). Like multiple regression analysis, AID studies the relationship between a dependent variable and a set of independent variables: however, unlike multiple regression analysis, it does not require a predetermined functional form. AID continually splits the data into subgroups on the basis of independent variables, but the "quality" of the subgroups is measured by the sum of squared deviations in the dependent variable. AID offers a number of advantages, among them being a well defined algorithm and no a priori assumption of a functional form. In addition,

it is well suited to large data sets when the number of independent variables is fairly small (since the number of potential branches on the AID "tree" is determined by the number of independent variables and the number of discrete intervals of each). Besides the restrictions of the number of variables, the major limitation for this study is AID's use of a dependent variable; as previously noted, outcome measures are not theoretically tractable (e.g., efficient costs would be exceedingly difficult to measure). Other limitations, suggested by Doyle and Fenwick (1975), include AID's tendency to inaccurately classify objects if the dependent variable is heavily skewed, instabilities in the resultant tree due to differences in the predetermined intervals for each in dependent variable, and ambiguities relating to the algorithm's stopping rules.

A comparison of six state-of-the-art classification methodologies is presented in Table 1.2. It is evident from this comparison that the most appropriate methodology which explicitly incorporates a measure of similarity between hospitals (i.e., explicit measure of a hospital pair's respective affinity), does not require assumptions regarding underlying distribution of the multivariate data, and most easily accommodates large scale data sets, is cluster analysis: the methodology of choice in this study.

III. Report Outline

The remainder of this report is organized as follows: Chapter Two develops a general economic framework for selecting classification variables. Empirical measures from the data base are selected to represent the variables in the classification process itself. The statistical methodology developed for the clustering problem is explained in Chapter Three. This chapter includes detailed discussions of similarity measure calculation, group determination, and validation and testing of the resultant clusters. The fourth chapter presents empirical findings from the data base of 1,070 randomly selected short-term general hospitals, including descriptions of the clusters found for both the complete sample and a subsample of 194 hospitals. The summary and conclusions are presented in Chapter Five, and listings of all computer programs are included in Appendix C.

TABLE 1.2

Comparison of Alternative Classification Procedures

CRITERIA	MULTIPLE CROSS-CLASSIFICATION	FACTOR ANALYSIS	REGRESSION ANALYSIS	DISCRIMINANT ANALYSIS	AID/AUTO GROUP	CLUSTER ANALYSIS
Multivariate Normal Assumption?	NO	YES	YES	YES	NO	NO
Linearity and/or Additivity?	NO	YES	YES	USUALLY	NO	NO
Dependent Variable Required?	NO	NO	YES	NO	YES	NO
Similarity Measurement Objective?	NO	NO	NO	YES	YES	YES
Predetermined Functional Form?	NO	YES	YES	YES	NO	NO
Predetermined Discrete Intervals for Variables?	YES	NO	NO	NO	YES	NO
Computationally Efficient?	NO	YES	YES	YES	USUALLY*	YES

* If the number of characteristics is small.

CHAPTER TWO

ECONOMIC RATIONALE AND VARIABLE SELECTION

I. Introduction

Chapter One outlined the various classification methodologies that are available and discussed their strengths and weaknesses from a statistical perspective. However, even the most appropriate methodology will yield misleading results unless the variables chosen for use in the classification process are also correct. That is, while the choice of methodology determines the path leading from the inputs to the conclusion, the conclusion itself is determined by the choice of inputs: the classification variables.

It is interesting to note that in spite of its importance, the issue of variable selection, unlike the issue of methodology selection, has received practically no attention in the literature. In none of the studies mentioned above was there a formal development of the variable selection criteria from a conceptual basis. That is the task of this chapter. A conceptual framework which suggests the variables that should be selected will be developed and applied to the problem and data set at hand.

II. Conceptual Framework

A. Grouping and Prospective Reimbursement

It was pointed out in the introduction that perhaps a large share of the cost increases of the last decade have stemmed from the fact that cost reimbursement relaxes the market constraint from the insured portion of the hospital's business. The objective of prospective reimbursement is to fill the void by approximating the constraints of a properly functioning market, thereby inducing hospitals to operate more efficiently both in the sense of technical efficiency and in the composition and nature of the output they produce.

Price is the constraint of the market. That is, firms face a price for the goods they produce: a price that is determined by aggregate market forces and reflects the cost of production. In competitive situations, prices are not subject to the influence of any single firm. Thus, the setting of an appropriate prospective rate should be analogous to the determination of an optimal market price.

The classification of hospitals into subgroups for the purpose of rate determination is merely recognition that, as in other industries, optimal market prices may differ across producers as market conditions differ. Thus, a single rate for all hospitals is not appropriate: prospective rates should differ across hospitals as market conditions differ. The objective of grouping is to identify the market conditions that would lead to price variation in a competitive market and to classify hospitals into groups based on their

similarity with respect to these criteria. The implication is then that cost variation observed within these groups arises from some other "unjustified" source and should not be recognized in the rate structure.[8]

The importance of these grouping criteria to the success of the reimbursement scheme cannot be over-emphasized. To avoid imposing unnecessary and inappropriate hardship on providers, account must be taken of differences in exogenous sources of cost variation. Failure to adjust for these differences will result in a system that is inequitable and arbitrary, resulting in serious distortions in the short run, and a movement of providers out of the industry in the long run. On the other hand, adjusting rates for cost differences arising from other sources allows "leakage" in the control mechanism (i.e., allows firms to escape the constraint) and may provide incentives for distorting the system further. There is substantial evidence from other industries that firms can and do adjust to controls by altering uncontrolled aspects of their operations. In this way, the effect of the constraint on the firm's ability to pursue its objectives is minimized. For example, the airline industry has responded to the imposition of price controls by changing an unconstrained characteristic of the output vector (service frills) that it offers for a given price. Public utilities, facing a ceiling on the rate of return to capital they may earn, have increased their capital stock in order to be allowed to earn more absolute profit dollars. (See Noll, 1973).

This list goes on. To avoid creating like incentives in the hospital industry, the selection of criterion variables used to group hospitals must receive as careful attention as the actual mechanism through which controls are imposed.

B. Cost Influencing Factors

Since prospective rates are to serve as producer prices, the appropriate place to turn for insight into the selection of proper grouping variables is the theory of the market. Economic theory suggests that in a world of perfectly competitive profit-maximizing firms, the prices of goods and services observed in any market reflect the opportunity cost of the resources used in producing those goods and services. Further, since firms are motivated by profits, their incentive is to produce exactly that combination of goods and services demanded by consumers, given tastes, incomes and prevailing prices for factors of production (which, given the assumptions of this section, represent marginal costs). In this situation all firms are driven through competition to a single price for a given output (adding the assumption of adequate information flow), and similar goods of different quality will exist at different prices only to the extent that consumer preferences dictate these differences (i.e., imported French wine will be purchased at higher prices than some domestics, but the market for gold lunchboxes is small).

Thus, prices (and costs) vary across firms because of output differences. However, since output has a number of dimensions, cost differences can arise

---

[8] It should be noted that variable definition and measurement problems reduce the accuracy of this proposition in practice.

when any aspect of output differs across firms. For example, even in an industry of single product firms, there may exist differences in the nature of the product, even though it is given only one name. That is, hospitals may have different "efficient" cost curves if one institution's appendectomy patients are diabetic and obese while another hospital provides appendectomy to otherwise healthy patients without complications.

Further, there may be quality differences leading to cost variations. If the wine that is aged for several years in wooden casks is of higher quality than a brew marketed from steel vats after one year, its higher costs will be covered by higher prices as long as consumers perceive (and are willing to pay for) the difference in quality. The appropriate analogy in the hospital industry might also focus on outcomes. If like patients with a given condition in hospital A have a lower rate of mortality and morbidity following admission than do those in hospital B, patients may be willing to pay a higher rate to be treated in the former. Some may also be willing to pay more for nicer surroundings, added comfort or more privacy.

In a multiple-product firm, the situation becomes more complicated. Ideally, variable costs are allocated to the various outputs and rates are set according to marginal production costs. Where this is the case, the preceding arguments hold. The marginal cost of producing a Volkswagen tune-up of a given quality should be the same across repair shops, regardless of the other services that may also be available in any shop (abstracting as before from exogenous factors). The same is true of the inpatient setting: a diagnostic chest x-ray given in a teaching institution should be no different nor more costly than the same procedure performed in another type of inpatient facility. The parallel in this instance is less exact, however, since the availability of some specialized equipment and personnel may affect the probability of experiencing a given outcome. That is, a delivery may take no more resources in a large hospital if the delivery is normal, but the outcome of a complicated birth might be enhanced by the presence of backup facilities (e.g., post-natal intensive care units, premature nurseries, etc.) unavailable in a smaller institution. Therefore, a woman who believed the probability of complications in her delivery to be large might well be willing to pay the higher rate of the larger hospital, even if the birth turned out to be normal, because of the reduction of risk. Over time, if these differences are indeed reflected in charges (and if patients pay their own charges), this suggests that women who expect normal deliveries will gravitate to the smaller (less costly) hospital, leaving the larger institution with a more complex mix of deliveries and an appropriately higher average daily cost.

Thus, price differences within a market in this setting are observed only to the extent there exist differences in the dimensions of output. Further, these output and quality differences reflect the exogenous preferences of consumers.

Price differences for a given kind and quality of output may exist across markets in this setting only if there are input market imperfections leading to long run differences in factor prices across markets, or local regulations that have an impact on cost (if lower priced--gross of shipping charges--goods from other markets cannot be transported in) (See Table 2.1).

TABLE 2.1:

Major Sources of Price Variation Among Competitive Firms

1. Nature of Output
2. Differences in Quality of Output
3. Output mix for Multiproduct Firms
4. Input Price Differences
5. Local Regulations
6. Scale Economies Across Markets

An important point to note is that in this setting characterized by perfect competition and adequate information flow, differences in market-generated prices spring only from forces exogenous to the individual firm. Price differences arising from differences in endogenous factors such as technology employed, age or size of plant, or input mix will not be sustained in the long run: firms that cling to their endogenous differences will be forced out of business by the lower cost competition.

C. Implications for the Hospital Industry

What does this theory imply for the determination of prospective rates in the hospital industry? First, one must question whether the assumptions that lead to the above conclusions hold in the market for hospital services. If so, then it follows that the level of cost in the industry is appropriate, the observed output mix is appropriate, and prospective rates reflecting these costs are appropriate. In this situation, hospitals should be grouped on the basis of observed cost; hospitals with dissimilar costs should be in different groups.

Unfortunately, the assumptions do not hold. While perfect competition may be an economist's fantasy, a strong argument could be made that even "adequate" competition (adequate in the sense of allowing a close approximation to the previously outlined conclusions) does not exist in most hospital markets. Barriers to entry in the form of capital availability, certificate of need laws, and AMA restrictions on the supply of an essential complementary good (physicians) act to limit the forces of competition. In addition, the presence of insurance on a wide scale further relaxes the ordinary market constraints. As the point of hospital service price approaches zero, the price sensitivity of consumers is greatly reduced, and demand increases. This situation is exacerbated by fee for service payment of physicians and cost reimbursement of inpatient facilities, both of which further encourage the expansion of output.

Further, information is costly in this industry. It was noted in Chapter One that the lack of knowledge about the product leaves the consumer dependent on the provider, facilitating the influence or generation of demand by physicians and/or hospitals. Further, information regarding prices and insurance coverage is not always readily available. These two facts increase the opportunity for non-competitive behavior among firms.

The implication of dropping the assumptions of perfect competition and free information flow is that market-generated prices can no longer be assumed to reflect "true" minimum costs since there may be a margin for monopoly rents. Without perfect competition, price differences reflective only of differing degrees of monopoly power may be sustained over time within one market (the example of monopolistic product differentiation is applicable here) or across markets. In this situation, taking market prices as optimal for use in a rate determination scheme is inappropriate.

Another problem is that most firms in this industry are legally non-profit entities. Thus, since hospitals may not be attempting to maximize profits, it is no longer clear as it is in other industries that their objectives will be furthered by producing the combination of services (output) most valued by consumers. Given the insurance-induced price insensitivity of consumers, firms have no valid signals as to consumer preferences. This fact, combined with the other barriers to perfect competition, means that the opportunity of the non-profit firm to pursue its own objectives without fear of financial losses is increased. If these objectives are furthered, for example, by expanding the production of one particular type of service or increasing the quality of output beyond the level indicated by consumer preferences, an inefficient output mix will be produced by the industry. This situation is exacerbated to the extent that providers can influence consumers' purchase decisions.

Neither is there the assurance of technical efficiency in the sense of producing the largest amount of services for a given set of inputs, an assurance guaranteed in a profit maximization setting, since the value to the non-profit firm of the profits lost through this type of inefficiency is likely to be small (unless profits may be accumulated for investment in projects that yield utility) (Clarkson, 1972). There is also the possibility, as suggested by some authors, that specific inputs rather than outputs enter the hospital's objective function (Lee, 1971). Thus, the expectation of technical efficiency is further reduced. Feldstein (1971) goes a step further to suggest the inclusion of selected input prices in the vector of objectives, again implying distortions in the input mix.

Thus, observed interfirm price and output differences in this largely non-profit market characterized by imperfect competition and provider influence of consumption decisions may not be assumed to reflect consumer preferences expressed in light of relative production costs. It may be more likely that they reflect (at least to some extent) differences in individual firm objectives or differences in the level of insurance coverage. Thus, since incentive reimbursement systems are superimposed on an industry that has been and continues to be heavily insured and primarily non-profit oriented, this result implies that rates set on the basis of historical cost and the firm's status quo may be inappropriate. This suggests that grouping hospitals according to observed costs and setting a separate rate schedule for each group (which probably amounts to separate rates for each hospital) will only perpetuate existing inefficiencies.

How then are appropriate groups to be determined? The previous examination of the perfectly competitive situation would indicate that hospitals should have different costs (and therefore face different prices or rates) only when they differ with respect to the exogenous characteristics outlined in that discussion: input prices, product mix (in multiple product firms), quality and nature of output.

This point is very important. Allowing a hospital to move to a group with higher rates because of a change in an endogenous factor encourages other hospitals to change this factor purposefully to increase their own rates. Therefore, an appropriate grouping scheme should be based on variables that cannot be manipulated by the institutions' administrators.

The task of the next section is to address more specifically the grouping variables that might be used in the hospital industry in light of the theoretical considerations.

III. The Variables

While the theory of the market presented in Section II provides some insight into the exogenous sources of cost differences among efficient firms, the translation of this theory into the specific variables that should be used to group hospitals for rate setting cannot in general be made without reference to other program parameters as outlined in Chapter One.

That is, the subset of appropriate grouping variables (within the set of variables identified by the foregoing theoretical considerations) will differ with the unit of payment that is chosen (e.g., total budget, per case--diagnostic specific or all-inclusive, per diem, or per service), and the costs that are to be included for control (e.g., total costs, routine costs, etc.).

The scope of the program in terms of its specific objectives will also influence the choice of grouping variables. This point relates to the discussion in Chapter One of the two types of inefficiencies that are thought to exist in the industry: technical inefficiency and product choice inefficiency. If the control program limits its focus to the first type, the grouping criteria are somewhat different than in a more ambitious program that attempts to control both sources of cost increase.

The impact of program objectives will be discussed first, followed by an examination of the grouping variables appropriate to the various program designs.

A. Program Objectives and Variable Selection

In Section II it was concluded that differences in the nature and quality of output are among the prime exogenous sources of cost variation among firms. However, it was also suggested that these variables may not be determined exogenously in the industry in question. That is, a consumer who does not have adequate information concerning, for instance, the impact of various

procedures on his health for a given condition must rely on the providers of health services to direct his medical purchases. This fact combined with the fact that the consumer likely faces below cost service prices because of insurance suggests that physicians and/or hospitals can direct the consumption of services both in number and in kind.

It is important to note that the extent to which "extra" or overly specialized services will be suggested by a provider depends in part upon the firm's objective function. If the firm is a pure profit maximizer, it will direct an increase in consumption of its most profitable services. A regulatory mechanism that is tight enough to eliminate economic profit from all services would effectively eliminate this distortion since nothing would be gained by such actions. However, if the firm is maximizing some non-profit objectives such as quantity of services or the production of a subset of services (e.g., highly specialized procedures), a break-even reimbursement rate will not curb the provider's incentive to encourage consumption.

Further, if the physician rather than the hospital is the prime source of advice for the consumer (the Feldstein "agent", see Feldstein, 1974), setting a break-even rate for the hospital will have no deterrent effect on over-consumption if the physician-agent is deriving some benefit, either financial or otherwise, from it. The hospital in this case is a passive actor.

These two points hold for every dimension of output. Including a quality measure in the set of grouping variables so that hospitals with higher quality are allowed higher rates encourages the production of "Cadillac" care, even when the value of the additional quality to the consumer is low relative to its true cost (e.g., all expectant mothers would be encouraged to deliver in an institution with a full range of back-up services--at higher cost--even if there was a very small probability of complications). The same argument holds with respect to plant size. While differences in scale of plant may be appropriate across market areas due to differences in the level of demand, adjusting for scale differences for all hospitals regardless of location will produce undesirable incentives for the output-maximizing hospital.[9] This point cannot be over-emphasized. Building such incentives into the reimbursement mechanism may result in a system that fosters higher rather than lower costs. Further, in the long run technology and output characteristics will be determined by the objectives of individual firms, encouraged by system incentives, rather than by the individual or collective preferences of consumers.

It is possible that even the composition of output (hereafter referred to as case mix) can be influenced by the provider, if in name only. That is, routine appendectomies might be registered as appendectomies with complications if producing more of the latter moved the institution to a group with a higher reimbursement rate (or carried a higher reimbursement rate directly).

---

[9] Most empirical studies indicate that scale differences have little impact on cost once account has been taken of differences in the nature of output. If some confidence could be placed in these studies, this would suggest that no further rate adjustment should be made on the basis of plant size (e.g., number of beds). However, the methodology and data used in most scale investigations has been sufficiently questionable so that their results must be regarded with some caution.

The important point to note is that the problem of regulating firms for which demand constraints are in general not binding (because of the presence of insurance) is much more complex than regulating monopolistic firms in other industries. The problems are exacerbated by the non-profit motive, but this is secondary in its importance to the insurance issue. In this situation, the regulator can choose between two basic objectives. The first is to accept the insurance-induced distortions, treat all dimensions of demand as exogenous, and attempt only to prevent large interfirm differences in technical inefficiency. The appropriate adjustment variables for diagnostic-specific service rates in this case are measures of product differentiation: the nature and quality of services; and the set of exogenous factors: factor prices, unionization, and local regulations.

The second alternative is to attempt to use incentive reimbursement as a mechanism not only for controlling technical inefficiency, but also to attempt to correct the product choice inefficiency brought about by insurance. This is a much more complicated task, since it implies knowledge of consumers' true demand curves. If such information were available, or if society were willing to specify acceptable levels of demand for various services (defined across all product dimensions) by socioeconomic and/or demographic characteristics of the population, appropriate areawide budgets could be developed based on these characteristics. The task of the local regulator would then be to accept bids from area providers for each type of service and allocate the area's budget to the different providers according to the lowest bid. Obviously, this type of approach is infeasible at the present time. A similar method which may be more feasible is to attack the problem from an altogether different angle. A restructuring of insurance coverage to focus on desired benefits without removing the price sensitivity of consumers would reduce (if not remove) the need for provider regulation.[10] Regulatory focus in this situation might revolve around provision of information and more traditional monopoly control.

However, since the latter alternatives are somewhat beyond current possibilities, it is desirable to seek a middle ground between complete passivity (i.e., treating all dimensions of demand as exogenous) and major insurance reform.

The approach which will be assumed in the remainder of this discussion is one such middle ground. Rather than treating all dimensions of output as exogenously determined, it is assumed that the mix of output (i.e., diagnostic case mix) and the nature of output (in terms of severity) are, if not purely exogenous,

---

[10] One possibility is to follow the lead of some European countries and issue medical care vouchers to consumers. Consumers are then responsible for choosing the source of care they receive as well as allocating the fixed sum in the manner that best reflects their preferences.

at least more difficult for hospitals and physicians to influence.[11] Quality differences are ignored with the assumption that providers accepted for Medicare (or more general) participation in the reimbursement program meet some minimum standard of quality. Since standards can always be adjusted, this level is taken as acceptable. The justification for this approach is twofold. First, the definition of quality is somewhat ambiguous. High quality treatment is not necessarily synonymous with highly specialized, technologically advanced care, although that correlation is too frequently drawn. That is, a small hospital concentrating mainly on simple procedures may produce higher quality care as measured in terms of outcome for its diagnostic mix than a hospital with very advanced equipment but sloppy administration and/or no expertise. Thus, high quality need not mean high cost (Shortell, 1976). The second point is that quality is highly endogenous. While there is much disagreement in the literature regarding the exact elements of the hospital's objective function, quality of care is often cited as a parameter over which hospitals exert much control and one which likely yields prestige to both the administration and medical staff.[12] Therefore, given that the (insured) market places little constraint on providers' actions, allowing upward rate adjustments for increased service quality is almost certain to encourage "Cadillac-ization" of care well beyond the desired level.

B. Program Design and Variable Selection.

Given the above assumptions about the scope and objectives of the selected reimbursement scheme, the grouping variables that are appropriate, in general, depend upon the control parameters of the program. There are, however, a number of cost-influencing variables identified in Section II as exogenous whose importance is independent of program design.

The first in this set of variables is the vector of input prices.[13] Failing to adjust for input price differences would penalize the institution located in an area, for example, of high wages and prices in a way that a market-determined price would not. That is, price differences reflecting factor price

---

[11] Grouping hospitals only on the basis of the demographic characteristics of the market in which they are located has a number of problems. First, the hospital's market is very difficult to specify. Frequently it does not conform to governmental boundaries (i.e., SMSA, county, etc.) for which such data are readily available, especially in the case of rural hospitals or specialized facilities that draw from a regional or sometimes national market. Further, even assuming these markets could be identified, adjusting only for the demographic characteristics of the population implies that all hospitals in an area should be identical. However, given the indivisibilities of much of the capital equipment used in the industry it is likely that institutional specialization is desirable and would lead to lower system costs.

[12] See, for example, Feldstein (1971); and Newhouse (1970).

[13] If more convincing evidence in support of the Feldstein proposition that hospitals influence nursing wages is found, this variable must be excluded.

differentials will be sustained in the market as long as the cost of transporting the output (in this case hospital services) is not zero. However, in adjusting for factor price differences, care must be taken not to dilute the firm's incentive to respond to factor price changes by substituting away from an input whose relative price has risen. For example, if a locality suffers an out-migration of RN's causing the wage of this type of personnel to rise, the institution whose reimbursement rate is adjusted to exactly offset this increase will have little reason to replace (where possible) some of its RN's with LPN's and other relatively less expensive personnel.[14] One possible means of handling this problem is to use a factor price index as the adjustment variable, rather than plugging in each individual input price. The development of such an index, however, requires some knowledge of the hospital's production function since the appropriate weight of each factor price in the index is the coefficient of the factor in the production relationship.

For the same reasons, local regulations and perhaps the extent of unionization of hospital personnel (which relates to input price levels) must also be recognized. Local regulations which restrict hospital operations may increase costs by preventing the hospital from choosing the least costly factor mix (e.g., certificate of need which restricts the use of capital). The impact of unionization will be partially reflected in wage levels. However, additional costs such as those arising from strikes (or costly negotiations to prevent strikes), differences in fringe benefits, or working conditions need to be recognized explicitly if it is felt that hospitals cannot (or should not) affect the extent of unionization of their employees. Again, no windfall losses or gains should accrue to providers because of this set of factors over which they exert little control.

Hospital costs will also be higher in rural areas where demand is not sufficient to intensively utilize indivisible pieces of hospital capital. Higher prices to cover the higher costs would be sustained in the market since, for some set of basic services, it is less costly for consumers to support an under-utilized facility than to travel to where services can be obtained or to do without. The problem of encouraging capital accumulation in rural hospitals beyond the optimal level dictated by consumer preferences is minimized if rates are set based on comparative rural hospital costs. Intra-group collusion to "game" the system is not likely with this subset of providers. Thus, a variable to account for this situation should be included in the set of grouping criteria.

---

[14] If prospective rates are paid as negotiated with no end of the year adjustment to reflect differences in negotiated and actual costs, the profit-maximizing hospital will respond appropriately to factor price changes even with built-in rate adjustments, since any cost savings arising from such actions will increase profits. In the non-profit sector, this will happen to the extent the hospital can use the savings in salaries for a more desired purpose (e.g., investment). If costs can be sorted out in such a way that only costs for identical products are being compared, variation should result only from factor price differences.

C. Program Design Parameters

The output variables reflect differences in the output characteristics assumed to be exogenous in the discussion at the beginning of this section. However, the particular variables chosen depend upon the level of aggregation of control and thus the program design parameters selected. Each possibility will be discussed separately. In all cases the set of variables discussed above as independent of program design should also be included.

C.1. Per Service Reimbursement

If providers are reimbursed separately for each service rendered, no output adjustment is necessary as long as hours of nursing service are billed separately also. This is true regardless of whether only routine costs are included or whether all non-teaching costs are controlled.[15] The only difference is in the number of rates that are required in the two situations.

Per service reimbursement allows (encourages) the maximum amount of provider-generated demand. Since the total budget increases as admissions, length of stay, and/or intensity of service (tests and services per day) increases, the output-maximizing hospital has the incentive to increase all three parameters. Quality conscious institutions are encouraged to increase the number of procedures performed per visit. Only the marginal profit-maximizing firm is given neutral incentives as long as the rate of each service is set equal to efficient production cost. Given the strong influence of physicians and hospitals on the output vector, it is unlikely that this payment mechanism would ever lead to lower hospital costs in the absence of very strong direct controls on output.

C.2. Per Diem Reimbursement

C.2.a. Routine Costs

The only appropriate output variables in the case of the more aggregate per diem reimbursement of routine costs (the current HCFA design) are a measure of case mix and case mix severity since these measures are likely to affect the optimum intensity of nursing services. Other output measures are inappropriate since they only affect costs that are outside the purview of the program.

The incentive facing hospitals under a per diem rate is to increase length of stay since later days generally are less service intensive and therefore less costly. That hospitals can in fact respond to a per diem control by keeping patients longer is implied by a study of prospective payment plan in downstate New York. Dowling, et. al. (1976) found some evidence that while per diem costs may have fallen slightly as a result of the control program, the increase in average patient stay probably wiped out any per case savings.

---

[15] Teaching costs will be discussed separately below.

Control is further eroded by including only routine costs in the reimbursement program. Hospitals can escape the constraint by increasing the number of ancillary services and/or artificially allocating some of the costs previously designated as "routine" to ancillary accounts. Given the state of the art in hospital accounting, such practices might be difficult to detect.

C.2.b. Total Non-Teaching Costs

When the control measure encompasses all non-teaching costs of the institution, the theoretical considerations of Section II dictate that additional variables to capture output differences are appropriate. In this situation, variables reflecting both case mix and case mix severity should be included in the grouping criteria since the composition of output determines the appropriate mix of ancillary services, and severity again influences both nursing intensity and ancillary services. The previous comments regarding the desirability of a per diem rate are applicable here as well.

C.3. Per Case Reimbursement: All-Inclusive Rates

Since average cost per case exactly equals average daily cost multiplied by average length of stay, differences among hospitals producing identical case mixes will occur only as there exist inter-institutional differences in treatment patterns or in length of stay. Since these differences are appropriate only as dictated by the case mix modifier (severity), including both case mix (where total costs are controlled) and case mix severity measures in the grouping criteria, eliminates the necessity for further grouping differences.

All-inclusive per case reimbursement of hospitals is possibly the most viable alternative. While this control mechanism creates an incentive to increase the number of admissions (cases treated), the hospital probably has less control over this variable than service intensity or length of stay. As noted below, a more aggregate unit of payment (e.g., total budget) requires an explicit quantity of output adjustment to avoid arbitrary treatment of institutions serving changing populations. Thus, increases in output (of a form determined by the type of adjustment used) may also be encouraged with a more aggregate control.

A further incentive created by this mechanism is for hospitals to attempt to change the mix of output that is produced. This is a problem created by any payment system based on averages. The firm is induced to increase production of the below average cost output and decrease production of the above average cost output. Thus, hospitals would attempt to encourage "easy" admissions (e.g., those whose projected cost was below the reimbursement rate), and refer away the "difficult" admissions (e.g., those whose projected cost was above the reimbursement rate). Anecdotal evidence suggests that this kind of substitution is possible, though data have not been available to formally test its magnitude.[16]

---

[16]The institution's ability to continue this substitution over time is limited by an annual determination of the reimbursement rate based on projections from the previous year's output mix for that hospital.

In addition to encouraging a substitution of one kind of output for another, this payment mechanism encourages hospitals to alter the characteristics of any given type of output. That is, the institution will attempt to use fewer resources to produce each admission, regardless of case type. While this is precisely the intention of the control (since, with factor prices constant, this is the only way costs can be reduced), it is argued that quality may suffer. However, given the bias toward high quality care on the part of physicians and the direct quality controls imposed elsewhere in the system (e.g., PSRO's, accreditation bodies, etc.), this is only likely to be a serious problem where rates are established well below current cost levels.

It seems clear at this point that the payment mechanism by itself cannot control all possible dimensions of hospital operations. Therefore, an effective cost-containment program almost certainly must embody a check on output as well as a per unit rate ceiling. Such a constraint on individual procedures or days of care would be difficult to design and enforce given the degree of endogeneity of these two variables as well as the fact that no accepted standard exists as to the number of procedures or days of care that are "appropriate" for a given diagnosis. Thus, a constraint aimed at either of these variables necessarily must involve statements about acceptable medical practice and therefore run the risk of being nebulous or arbitrary. A constraint on total admissions has two advantages: it need involve no statements about acceptable medical practice, and as noted above, the number of admissions is probably less subject to manipulation by the hospital than are the other two variables.

C.4. Per Case Reimbursement: Diagnostic-Specific Rates

An alternative to the all inclusive per case payment unit (e.g., average cost per case reimbursement) is the determination of diagnostic-specific per case rates. Thus, instead of setting one rate for all admissions and grouping by the exogenous variables (factor prices, etc.), case mix, and case mix severity, separate rates could be set for each different case-type (diagnosis). In this situation, the appropriate grouping variables are again the exogenous factors (input prices, etc.) and case mix severity. If the diagnosis definition includes a severity modifier (e.g., if acute appendicitus complicated by diabetes and uncomplicated appendicitus are defined as different case types, only the exogenous variables need appear in the set of grouping criteria.[17]

An appealing feature of the diagnostic-specific per case payment system is that the problem of output substitution noted above for all-inclusive case rates is reduced or eliminated, depending upon the specificity of case-type definitions and the accuracy of the case-specific rate (obviously if the payment rate for routine appendectomies is set well above its expected cost of production hospitals will attempt to do more appendectomies). The issue of quality reduction as the characteristics of each type of output changes remains.

Another advantage of this system is that no case mix variable need be included in the set of grouping criteria since differences in case mix across hospitals are accounted for explicitly in the number of payments of each type the

---

[17] This is essentially the Worthington approach (See Worthington and Hixson 1975).

hospital receives. Thus, case mix data need not be collected by the regulatory body for purposes of grouping. These data become available, however, as institutions request payment, just as information regarding the number of days of care provided is made available under current per diem systems.[18] On the other hand, the fact that hospitals supply this information in the form of payment requests leaves an uncomfortable degree of room for data manipulation. That is, if diagnosis A commands a higher rate than diagnosis B (which involves basically the same body system and requires many of the same procedures), there is a large incentive for the hospital to label all diagnostic B patients as patients with diagnosis A. The ability of the regulatory agency to constrain this type of action is extremely limited, especially given that there is often legitimate disagreement among physicians over questions of appropriate diagnosis.

A further problem with this approach is that separating a hospital's total costs into case-specific costs with any degree of accuracy is likely to be extremely difficult given the current accounting parctices of hospitals. Thus, a diagnostic-specific per case system, while conceptually very close to an all-inclusive per case control, is probably a less desirable approach from a practical standpoint.

C.5. Total Budget Reimbursement

C.5.a. Routine Costs

This is the most aggregate payment unit. Where a total budget for routine costs is determined prospectively, account must be taken of the quantity of output (i.e., days of care) produced as well as severity, as before. If two facilities have identical (severity adjusted) per diem or per case routine costs, they will still have different total yearly general service costs if they produce different levels of output. Unfortunately, this is difficult since determining budgets for hospitals grouped by quantity of output encourages the output maximizer to expand even if all budgets are set at a break-even level. This distortion would have to be checked with an additional constraint. The earlier comments regarding the possibility of escaping control by allocating routine costs to unregulated accounts again hold.

C.5.b. Total Non-Teaching Costs

The above argument holds for total non-teaching budget reimbursement with the addition of a case mix variable. Again, some additional constraint is called for to discourage expansion beyond the optimal level (See Table 2.2).

---

[18] Pre-control baseline data, which would be extremely valuable in monitoring output changes resulting from the payment program, would still have to be collected directly.

TABLE 2.2

Program Design and Variable Selection

ALL SYSTEMS

1. Input Prices
2. Local Regulations
3. Extent of Unionization
4. Rural Markets

I. PER SERVICE REIMBURSEMENT

A. Routine Costs
   1. No Further Adjustment
B. Total Costs
   1. No Further Adjustment

--Increase Admissions
--Increase Length of Stay
--Increase Intensity of Service

II. PER DIEM REIMBURSEMENT

A. Routine Costs
   1. Case Mix Severity
B. Total Costs
   1. Case Mix Severity
   2. Case Mix

--Increase Admissions
--Increase Length of Stay

III. PER CASE REIMBURSEMENT: ALL INCLUSIVE RATES

A. Routine Costs
   1. Case Mix Severity
B. Total Costs
   1. Case Mix Severity
   2. Case Mix

--Increase Admissions

IV. PER CASE REIMBURSEMENT: DIAGNOSTIC SPECIFIC RATES

A. Routine Costs
   1. Case Mix Severity
B. Total Costs
   1. Case Mix Severity

--Increase Admissions

V. TOTAL BUDGET REIMBURSEMENT

A. Routine Costs
   1. Case Mix Severity
   2. Quantity of Output
B. Total Costs
   1. Case Mix Severity
   2. Case Mix
   3. Quantity of Output

D. Teaching Costs

To this point, no explicit mention has been made of teaching expenses. Teaching is an aspect that has often been identified as having a significant impact on costs (Lave and Lave, 1970); however, little study has been directed at the explanation of this observation. If teaching facilities serve a patient group with more complex conditions (as is sometimes argued) this would be accounted for via the case mix and severity adjustments. If the quality of care is higher in these hospitals, this aspect would also be addressed. Thus, only the issue of costs associated directly with training of new health personnel is omitted.

There are two arguments which might be made here. The first is that teaching costs might be accounted for in terms of lump sum program costs rather than a percentage increase in lab costs, routine service costs, CCU costs, etc. That is, the time resources required of the chief surgeon to instruct interns in his (her) department are appropriately charged to teaching rather than to surgery since it is a separate activity.[19] Second, since the benefits of the teaching program accrue not just to the patients in that particular institution but more generally to the population as a whole (as with medical research), there is no reason to argue that only those served in the teaching hospital should pay all associated teaching costs. A more appropriate funding mechanism might be based on general tax support, either from state funds or federal funds, depending upon expectations about migration.

Building instructional costs into the rate structure not only raises equity questions, but also has implications about optimal program size. If medical education is supported through lump sum grants, the amount of resources devoted to this activity is directly controlled by public policy at the appropriate level. When teaching costs are met with inpatient revenue, program size and composition (in terms of specialty mix) is determined by individual hospitals, and is influenced to some extent by demand for that institutions's services. Creating a rate control system that explicitly allows for teaching costs, especially as a percentage of total patient costs, encourages the expansion of teaching without tying its growth to increases in demand for personnel. However, it is possible that direct payment for teaching costs may not be feasible either practically or politically at this time. If this is the case, and if such additional expenses are judged appropriate at their current levels, a teaching variable could be included in the set of grouping variables.

E. The Problem of Weighting

Once the appropriate set of grouping variables has been identified, there remains the question of relative weights. That is, if input price and (exogenously determined) output differences are both "legitimate" sources of cost variation among firms, which is more important in terms of its impact on cost? If hospital A faces higher input prices, but hospital B produces a more complex output, which hospital would be expected to have higher costs, assuming homogeneity along all other dimensions including efficiency? The question can also be asked of the various kinds of output--e.g., brain surgery vs. hernia repair.

---

[19] There is a practical problem with this approach since teaching costs would be very difficult to identify given current hospital accounting systems.

There is no theoretical answer to the question. Empirically, it is possible to determine the relative share of each cost component, but this analysis will only produce appropriate shares if firms are producing efficiently in the economic sense. That is, regressing cost on factor prices and output characteristics will only yield coefficients that describe current hospital operations rather than those that identify the efficient cost relationship. Further, since it is possible (likely) that hospitals in different groups have different cost structures (i.e., there are slope as well as intercept differences in their respective cost functions), appropriate weights might differ across groups. Although these empirical approximations are admittedly imperfect, they may represent the most acceptable alternative. Another alternative, of course, is to weight each of the variables equally.

F. Partial Coverage

Regardless of the set of variables for which rate adjustments are made, the program will impact differently on the industry depending upon the subgroup of patients to whom the program pertains. If only costs associated with Medicare patients are controlled, the net impact on total institutional cost is likely to be small. Further, the hospital has at least two options for circumventing the system. The first is to allocate costs that are not covered by Medicare reimbursement to other patient groups, either directly to justify higher cost-based reimbursement or by charging prices in excess of costs to make up the difference.[20] Medicare costs are still lowered in this situation, but this decrease will be more than offset by increases in non-Medicare costs if the juggling requires any administrative resources. There are also problems of equity since non-Medicare patients are thus required to partially subsidize the services of the older group.

The second option open to the hospital is to refuse to admit Medicare patients. If this option were exercised, the impact could range from inconvenience to patients whose admissions were delayed because of transfers to a reversal of Medicare's role in access to care.

Slippage of this sort can be somewhat mitigated by building positive as well as negative incentives into the system. If rates are set as group averages with providers grouped according to the variables identified above, hospitals whose costs fall below the mean will make a profit on each Medicare admission. Thus, over time, Medicare patients may be shifted from high cost (above the mean) hospitals to low cost hospitals. If this is a pure substitution of one group of patients for another, total system costs will be unaffected. Medicare costs will be lowered only to the extent high cost hospitals drop Medicare participation completely so that their average costs are no longer calculated into the mean for purposes of rate determination.

IV. Variable Measures

The previous sections have identified the set of variables which, from a conceptual viewpoint, should form the basis of a hospital classification system.

---

[20] The extent to which this is feasible depends on the method of payment for non-Medicare services and the size of the Medicare group relative to the total patient load.

However, to translate the conceptual notion into a workable empirical framework for the purpose of implementation, it is necessary to identify specific measures by which the variables can be represented.

Ideally, the measures chosen are exact empirical reflections of the variables: a family purchasing power variable translates into median family income[21] for the calendar year under consideration, including wages, interest, dividends, and other factor payments; transfer payments (e.g., social security, welfare, etc.); income in kind (including imputed rental value of housing and employer contributions to fringe benefits); minus tax liabilities at all levels. Usually this translation of variables into their measures does not contain the complete conceptual thrust of the original variables. Data restrictions to a large degree constrain the set of measures available. In the previous example, data on in-kind income (especially imputed figures) are not generally available. Further, the ideal measure is often not fully specified by the conceptual model. That is, should windfall gains that accrue during the study year be included in the previous example since they add to consumers' purchasing power, or should they be omitted on the assumption that unexpected income is quickly tucked into savings and ignored by the consumer for the purposes of current desisions? Questions such as these are answerable only empirically, since a sound conceptual model could be developed to predict either outsome. Therefore, it is not always possible to derive a single set of ideal measures to reflect the appropriate variables even without data restrictions.

The conceptual framework identified six variables as sources of expected cost variation among efficient institutions, based upon the application of market price theory to the hospital setting: factor prices, unionization, external regulation, rural scale diseconomies, case mix composition, and case mix severity. The remainder of this section will discuss how each of these variables might ideally be represented empirically, and given the data restrictions of the present study, how they will be represented for purposes of this analysis. This information is summarized in Table 2.3. Finally, some attempt will be made to assess the expected impact on the resulting classification due to the substitution of available measures for ideal measures.

A. Factor Prices

The factor price variable is intended to pick up exogenous differences in the prices that hospitals in different input market areas must pay for these inputs. The ideal measure of this variable is sensitive not only to absolute price level differences, but also to the impact of selective differences. That is, a ten percent differential in the price of labor between two areas will have a larger effect on hospital cost than will a ten percent difference in the price of laundry detergent. Further, it must be recognized that the extent to which input price differences will be reflected in final costs will depend upon the substitutibility of inputs in the production process. That is, if labor

---

[21] Normally, median family income is used instead of average family income so that the extreme values of income may not distort the statistics.

TABLE 2.3

Empirical Measures for Conceptual Variables

1. INPUT PRICES
   - A. Hospital Wages
   - B. Manufacturing Hourly Wage
   - C. Transportation and Public Utilities Hourly Wage
   - D. Retail Hourly Wage

2. LOCAL REGULATIONS
   - A. No Measures Available

3. EXTENT OF UNIONIZATION
   - A. No Measures Available

4. RURAL MARKETS
   - A. Uniform Pressure Occupancy Index

5a. CASE MIX: ENDOGENOUS APPROACH
   - A. Number of Basic Services
   - B. Number of Quality Enhancing Services
   - C. Number of Complex Services
   - D. Number of Community Services
   - E. Number of Births/Number of Discharges
   - F. Number of Surgical Operations/Number of Discharges
   - G. Number of Outpatient Visits/Number of Discharges

5b. CASE MIX: EXOGENOUS APPROACH
   - A. Median Family Income
   - B. Percentage of Population Female and Aged 15-44 Years
   - C. Percentage of Population Aged 0-5 years.
   - D. Percentage of Families Earning Income Less Than $4,000
   - E. Labor Force Participation Rate Age 16 and Over
   - F. OB-GYN's per 10,000 Population
   - G. Primary Care M.D.'s per 10,000 Population
   - H. Percentage of Population non-White
   - I. Disability Rate Age 16-64 Years
   - J. Percentage of Population Aged 65 Years and Over
   - K. Percentage of M.D.'s Aged 60 and Over
   - L. Medical Specialists per 10,000 Population
   - M. Other Direct Care Specialists per 10,000 Population
   - N. Other Specialists per 10,000 Population
   - O. Surgical Specialists per 10,000 Population

6a. CASE MIX SEVERITY: ENDOGENOUS APPROACH
   - A. Percentage of Population Aged 0-5 Years
   - B. Percentage of FAmilies Earning Income Less Than $4,000
   - C. Percentage of Population Aged 65 Years and Over

6b. CASE MIX SEVERITY: EXOGENOUS APPROACH
   - A. No Further Measures Necessary

initially accounts for 50 percent of total expenses per unit of output, a 10 percent rise in the price of labor will only result in a 5 percent increase in the cost of a marginal unit if other inputs cannot be substituted (at least to some extent) for the now more expensive input. Thus, the ideal measure of the impact of factor price differences on output cost is an index formed as the weighted sum of relevant factor prices where weights reflect the input's coefficient in the production function.[22] Since the data to construct such an ideal measure are not available in the current study, a much less exact measure will be used. Factor prices will be represented as a vector with four elements: a computed average HOSPITAL WAGE, MANUFACTURING HOURLY WAGE, TRANSPORTATION AND PUBLIC UTILITIES HOURLY WAGE, and RETAIL TRADE HOURLY WAGE. The first, average hospital wage, will be constructed as the endogenous measure, PAYROLL FOR ALL OTHER PERSONNEL, divided by the sum of PERSONNEL--FULL TIME ALL OTHER REGISTERED NURSES, PERSONNEL--FULL TIME LPN's, and PERSONNEL--FULL TIME ALL OTHER (collected from the AHA data). Because this measure uses the hospital's own payments per employee it is only useful as an historic measure. That is, in an on-going system, the use of such an endogenous variable might encourage hsopitals to increase their payroll in the hopes of moving themselves to a more renumerative group.[23] The other three wage categories are used as proxies for the general cost of living in the area which primarily reflects prices of non-labor inputs (e.g., food, etc.).

---

[22] If there are two inputs, L and K (with prices, w and c respectively) used to produce some output, Q, then total cost is given by:

$$TC = wL + cK$$

Differentiating with respect to Q gives:

$$MC = \frac{\partial TC}{\partial Q} = w\frac{\partial L}{\partial Q} + c\frac{\partial K}{\partial Q}$$

The change in the cost of a marginal unit of output as one factor price, w, changes by one unit is:

$$\frac{\partial MC}{\partial w} = \frac{\partial L}{\partial Q} \qquad \left(\text{Similarly, } \frac{\partial MC}{\partial c} = \frac{\partial K}{\partial Q}\right)$$

Thus, the impact on marginal cost of finite charges in all input prices is:

$$dMC = dw\,\frac{\partial L}{\partial Q} \quad dc\,\frac{\partial K}{\partial Q}$$

But, $\frac{\partial L}{\partial Q}$ and $\frac{\partial K}{\partial Q}$ are simply the inverses of the marginal products of those inputs, or the inverses of their coefficients in the production function.

[23] In addition, the measure only captures a subset of factor prices and will be influenced by differences in the mix of labor in the categories used as well as differences in the wages of these markets.

The bias resulting from the use of this measure rather than the ideal counterpart is uncertain in direction and magnitude.

B. Unionization

The purpose of this variable is to capture the cost impact of unionization that is not reflected in wage levels. That is, management may incur additional operating costs because of the presence or threatened presence of labor unions: higher fringe benefits, costs associated with union bargaining, etc. The ideal measure of this variable includes the percentage of an institution's employees that are unionized (to capture the extent of power of the unions), the number of unions represented (as a measure of the transaction costs involved with union dealings), and some measure of the threat of unionization--perhaps the percentage of service employees in the county which are unionized.

In the present study, no data regarding unionization were available. It is unlikely that a serious distortion is caused by this omission.

C. External Regulation

The impact of external regulation on costs is not dissimilar in nature to that of unionization. Regulation restricts management decision-making thereby increasing operating costs (unless the regulation is not binding). Further, there are costs associated with dealing with the regulatory agency (i.e., filling out forms, appearing at hearings, etc.). As with unionization, the ideal measure of this variable captures both the direct costs of the restrictions and the "haggle" cost. The latter might be measured by the number of different agencies with which the hospital had to deal (although this is an imperfect measure since the extent of the administrative requirements of different agencies is likely to vary substantially). The former is more difficult to measure. Ideally, some attempt would be made to assess the total cost impact of various restrictions. For example, certificate of need legislation may induce the hospital to substitute labor or uncontrolled equipment for additional beds, resulting in higher average costs.[24] On the other hand, capital controls may act to limit entry (and therefore competition) thus decreasing the amount of resources that must be spent attracting physicians from competing institutions. Further, it must be noted that such cost assessments must be made from an empirical rather than a theoretical perspective since the implementation of a law often differs substantially from its legislative intent.

Again, no regulation data are available in the present study. The degree of bias resulting from this distortion cannot be known without some idea of the magnitude of costs arising from regulatory restrictions, and the variability of regulations across hospitals.

---

[24] A study done by Salkever and Bice (1976) lends some empirical support to this notion.

D. Rural Markets Variable

While cost differences arising from economies of scale are not generally characterized as justifiable for reimbursement purposes, in the special case of rural hospitals an exception is made. The argument has been made that rural communities may not be able to support a set of basic facilities at optimal capacity. The resulting higher average costs in this case are justifiable, however, since it may be less costly for the community to finance this excess capacity than to seek basic services elsewhere or to do without them.

While the general conceptual notion is clear, its translation into an empirical measure is much less so. The variable is attempting to allow for the impact of desired (by the community) excess capacity on costs. In a competitive situation, the appropriate adjustment would be found in the market as the amount consumers are willing to pay over what they would pay in a larger neighboring community (this is obviously influenced by the costs of getting to the larger town). Within the confines of a reimbursement system, however, the accurate measurement of this variable is necessarily imperfect.

The rural markets variable proposed for this study, referred to as the uniform pressure occupancy index (UPOI), is based upon the concept of uniform probability of overflow suggested by Rosenthal (1964). In brief, Rosenthal's approach computes those values of average daily census for each geographical area such that the probability of hospitals being filled to capacity will be the same irrespective of their size and location. Rosenthal's measure, however, is based on the average size of hospitals in a county--an assumption we tested vis-a-vis regression analysis, using county population, county density, and a dummy variable indicating SMSA/non-SMSA as independent variables. Two regression equations, one using the total number of county hospital beds and the other using the average number of county hospital beds as dependent variables, were determined from our complete sample of 1,070 hospitals. The results are presented in Table 2.4; from this table it is evident that using total county hospital beds is significantly preferred as an urban/rural indicator. Therefore, our development of the rural variable is based on the total number of hospital beds in a county (B) and differs from Rosenthal's measure in this respect.

The computations are also based upon the additional assumptions:

1. The probability of overflow for any hospital is constant for any given day.

2. The daily census is Poisson distributed.

3. The probability of overflow cannot exceed 0.01 (i.e., the overflow will not occur more than one day in 100).

The uniform pressure occupancy index is calculated by finding the value of the average daily census (ADC) for each county such that the probability of demand exceeding the total number of beds in that county is less than 0.01, and dividing ADC by the total number of beds in the county (B). For example, if the total number of county beds is 75, then, by the Poisson assumption, the following formula will compute ADC such that 75 total beds will meet the demand

99 percent of the time:

$$0.99 = \sum_{x=0}^{75} \frac{e^{-ADC}(ADC)^x}{x!} \quad . \tag{2.1}$$

Using a commulative Poisson table to solve the above equation, one finds that ADC = 55, and, by definition,

$$\text{Uniform Pressure Occupancy Index} \quad (UPOI) = \frac{ADC}{B} = \frac{55}{75} = 0.733.$$

While the calculation of UPOI is theoretically straightforward, a major obstacle is provided by the fact that cummulative Poisson tables do not readily exist for values of ADC > 30. However, using the fact that the Poisson distribution approaches the normal distribution for values of ADC > 30, equation (2.1) can be rewritten as:

$$\lim_{ADC\to\infty} \sum_{x=0}^{B} \frac{e^{-ADC}(ADC)^x}{x!} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{B} e^{\frac{-(x - ADC)^2}{2\sigma}} dx = 0.99 \tag{2.2}$$

where $\sigma = \sqrt{ADC}$ (Equation (2.2) states that x is normally distributed with a mean equal to ADC and standard deviation equal to $\sqrt{ADC}$). To find the value of B such that the area under the normal curve from $-\infty$ to B is equal to 0.99, let Z be a standard normal variate, such that

$$Pr\left[Z = \frac{B - ADC}{\sqrt{ADC}}\right] = 0.99.$$

The value of Z is then found to equal K from a table of standard normal values (where K - 2.33 for a probability of 0.99). In general,

$$\frac{B - ADC}{\sqrt{ADC}} = K \Rightarrow ADC = \left[\frac{-K + \sqrt{K^2 + 4B}}{2}\right]^2$$

and

$$UPOI = \frac{ADC}{B} = \frac{\left[\frac{-K + \sqrt{K^2 + 4B}}{2}\right]^2}{B} = \frac{K^2 - 2K\sqrt{K^2 + 4B} + k^2 + 4B}{4B} \quad .$$

Solving for UPOI, one finds

$$UPOI = 1 + \frac{K^2}{2B} - \sqrt{\frac{K^4}{4B^2} + \frac{K^2}{B}} \quad .$$

Since $\frac{K^4}{4B^2} \to 0$ for typical values of K (<3) and B (>40),

$$\sqrt{\frac{K^4}{4B^2} + \frac{K^2}{B}} \to \sqrt{\frac{K^4}{4B^2}} + \sqrt{\frac{K^2}{B}}$$

and the expression for UPOI can be written as follows:

$$\text{UPOI} \overset{\sim}{=} 1 - \frac{K^2}{2B} - \sqrt{\frac{K^4}{4B^2}} + \sqrt{\frac{K^2}{B}} = 1 - \frac{K^2}{2B} - \frac{K^2}{2B} - \frac{K}{\sqrt{B}}$$

$$\overset{\sim}{=} 1 - \frac{K}{\sqrt{B}} \qquad (2.3)$$

In the above example where ADC = 55, it was found that UPOI = 0.733 using the Poisson tables; using (2.3), UPOI = $1 - \frac{2.33}{\sqrt{75}} = 0.731$. For K = 2.33 (i.e., the probability of 0.99), typical values of the UPOI calculated from (2.3) are shown in Figure 2.1.

TABLE 2.4

Regression Analysis: Rural Markets Variable

Regression 1: Dependent Variable - Total Number of County Hospital Beds

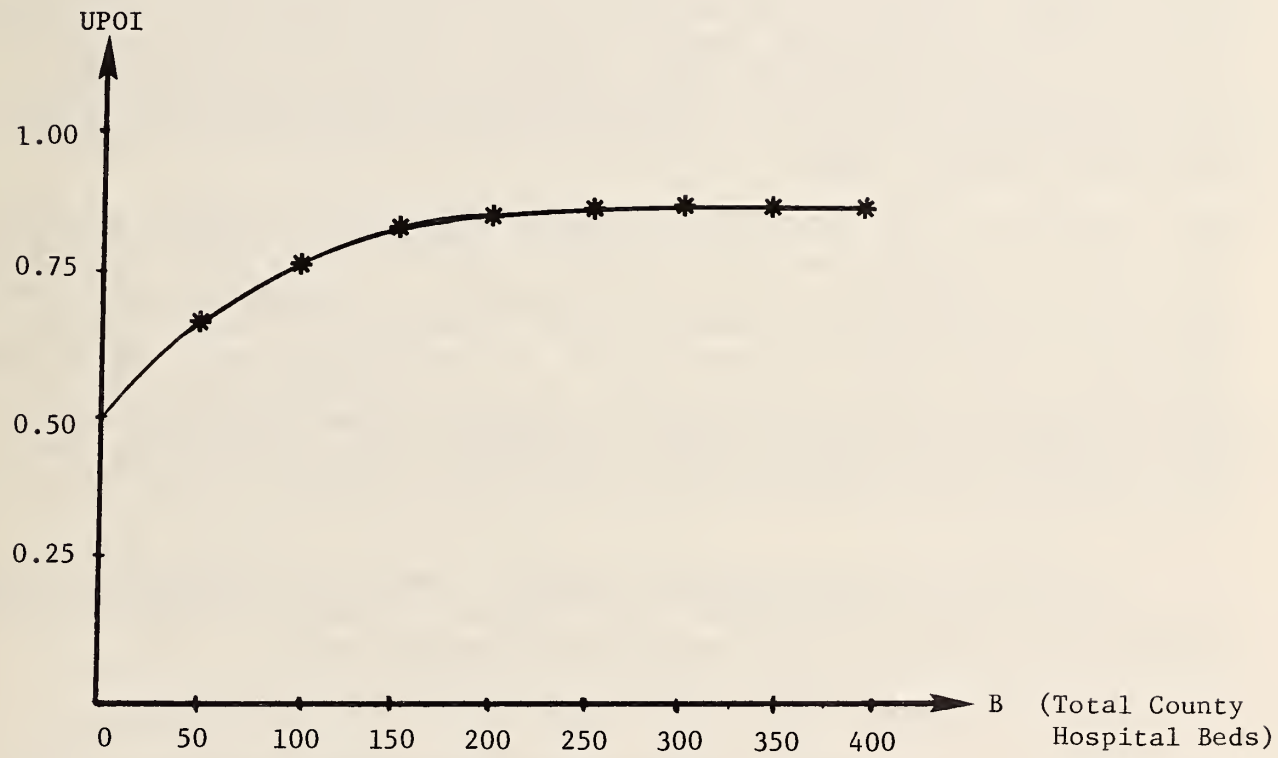
Independent Variable	Coefficient	Cummulative R^2	Overall F	Overall Significance
(Constant)	31.2	----	1703.5	0.0
Population	.0043	.871		
Density	.099	.892		
SMSA	78.35	.892		

Regression 2: Dependent Variable - Average Number of County Hospital Beds

Independent Variable	Coefficient	Cummulative R^2	Overall F	Overall Significance
(Constant)	104.2	----	8.1	0.0
Population	$.625 \times 10^{-5}$	.079		
Density	.0039	.094		
SMSA	104.67	.299		

FIGURE 2.1

Uniform Pressure Occupancy Index (UPOI)



E. Case Mix Composition

The ideal measure of this variable captures the range of product mix in a hospital. Since it is believed that product mix differences lead to average cost differences the description of case mix composition must be detailed enough to differentiate between two product types whose production costs differ. Ideally, every different product would be identified even if at present their cost differences are small since over time, changes in technology might convert these cost differences to ones of greater magnitude. Thus, case mix aggregation along departmental or disease code lines may not be appropriate if the product specification is expected to the useful in the long run.[25]

Many studies have used a scope of services index as a proxy for case mix.[26] The idea is that a hospital, for example, can only treat surgical patients if it has an operating room. While this fact is obviously true, the reverse correlation is not as strong, expecially in a non-profit oriented industry. That is, the presence of a coronary care unit of a given size in two different hospitals does not imply that the CCU occupancy rate is the same in both institutions, or that the number of cases treated in that unit as a percentage of total admissions is the same. Even the expressed desire to add a CCU to an existing structure need not, in a non-profit firm, mean that community demand for such a service is high.

More importantly, even if a scope of services index is an acceptable proxy for case mix in a retrospective study of hospital costs,[27] it is definitely not a good case mix substitute in a rate setting system. If the program allows hospitals with a more complex or wider range of services to be placed in groups with higher rates, the clear incentive for the administrator is to add services regardless of how intensively they might be used. The result over time would be a proliferation of underutilized facilities with substantial implications for both cost and quality of care[28] (see Glasgow, et. al., 1976). The correlation between case mix and the services index, even if perfect in the initial

---

[25] Obviously there are practical tradeoffs. Empirical investigations might indicate that aggregation to some level loses less in terms of accuracy than it might save in terms of additional data collection costs.

[26] See, for example, Berry (1974).

[27] Some on-going research at the University of Washington indicates that even retrospectively a fairly detailed scope of services index correlates poorly with case mix (the correlation coefficient is 0.03). This result must be interpreted with caution, as no adjustments have been made for case mix severity.

[28] Empirical evidence suggests that outside controls on capital expansion have been less than effective. See David Salkever and Tom Bice (1975).

period, would be small in future periods. Further, since data regarding scope of services (beyond the very simplistic AHA annual questionnaire data) would also have to be collected if it were to be used as a grouping variable, it would seem much more prudent to go after the real variable, case mix, rather than its proxy.

E.1. Exogenous Approach

For the present study, no explicit case mix data at any level of detail are available. Therefore, case mix composition will be estimated from two opposite approaches, the first, assuming a fixed set of hospital facilities (supply) in the short run, observed case mix as a function of exogenous and/or predetermined demand variables, such as population age, income, insurance coverage, etc. The problem with this approach is that while groups are being formed with individual institutions, demand variables are available only on a county-wide basis. The result of that data constraint is to interject a strong geographic bias into the system. That is, all hospitals located in the same county will have identical values for all of the grouping variables and will thus be grouped together. What is true, however, is that a hospital's market area is not necessarily equivalent to county boundaries. Specialty hospitals may draw from a regional or even national market (although a subset of that population) while local general hospitals may serve less than the entire county. Unfortunately, these catchment areas are difficult to define and appropriate data corresponding to their boundaries are not available.

The specific measures to be used in this approach are the following:

MEDIAN FAMILY INCOME

(CENSUS 1970 FEMALE POPULATION AGE 15-24 + CENSUS 1970 FEMALE POPULATION AGE 25-35 + CENSUS 1970 FEMALE POPULATION AGE 35-44) ÷ POPULATION SIZE

PERCENTAGE OF POPULATION AGE 0-5 1970

PERCENTAGE OF FAMILIES, INCOME LESS THAN $4,000

LABOR FORCE PARTICIPATION RATE AGE 16+

OB-GYN's PER 10,000 POPULATION

PRIMARY CARE MD's PER 10,000 POPULATION

PERCENTAGE OF POPULATION NON-WHITE

DISABILITY RATE AGES 16-64 (%)

(PERCENTAGE OF POPULATION AGE 70 AND OVER + PERCENTAGE OF POPULATION AGE 65-69)

PERCENTAGE OF POPULATION AGE 60-64

MEDICAL SPECIALISTS PER 10,000 POPULATION

OTHER DIRECT CARE SPECIALISTS PER 10,000 POPULATION

OTHER SPECIALISTS PER 10,000 POPULATION

SURGICAL SPECIALISTS PER 10,000 POPULATION

E.2. Endogenous Approach

Whereas the first approach focused exclusively on measures exogenous to the individual hospital, the second approach considers only endogenous measures. This approach assumes that case mix composition is accurately reflected by the facilities and services available in the institution. As noted in earlier sections, this assumption is possibly not justified in the short run and almost certainly not justified in the long run. In essence, this approach allows too much distinction among hospitals whereas the first approach allowed too little.

What is desired in this approach is to combine the available information on presence or absence of various facilities and services in such a way as to provide a meaningful description of the specialized asset composition of the institution. Using 46 dummy variables (corresponding to the 46 facilities and services which data are available--variables 124-169) is neither practical nor particularly enlightening. However, the literature provides no consensus on the appropriate method of combining the information into fewer, more useful measures (e.g., a facilities index). Here again the issue of variable weighting arises. Since different facilities, in general, have different implications in terms of their impact on institutional cost, it would be inappropriate to weight them equally in a system designed to focus on cost homogeneity. This has, however, been the approch taken in the literature, so that no tested weighting scheme is available for facilities and services. Therefore, it is proposed here to use a modification of the approach suggested by Ralph Berry (1973). Berry observes that empirically, hospitals can be classified into five groups based on the subset of facilities and services offered. In the present study, the endogenous measure of case mix for each hospital will be given as the number of facilities and services it has in each of Berry's four[29] categories. Thus, the measure is a vector with four elements.

Again, it is important to note that neither of the approaches used here are appropriate for use in an on-going control system. They are used here only for reasons of data availability.

F. Case Mix Severity

The ideal measure of this variable is something of an empirical question. That is, there is no conceptual argument pointing to the use of age, pre-existing condition, income level, etc. as a measure of case mix severity. The proper

---

[29] Only the first four of the five Berry groups (Basic, Quality Enhancing, Complex, Community, and Special) will be considered, since the last group termed "special" contains services such as chapel, hospital auxiliary, chaplainary, etc. which will have almost no effect on the case mix of the hospital.

measures are identified through the continual observation that (for example) older or diabetic patients require more resources to achieve the same outcome for any given condition holding constant other variables (e.g., price insurance, etc.) that might affect resource utilization. There are two difficulties. One is that some severity modifiers (e.g., obesity) might be appropriate only for some conditions (e.g., abdominal surgery) and not for others (e.g., broken arm). The second is that since the modifiers must be identified empirically, it may be difficult to distinguish between "true" measures of severity and taste variables.

Lacking any conclusive evidence regarding the appropriate set of severity modifiers and recognizing data constraints, the present study will use the following measures:

PERCENTAGE OF POPULATION 0-5 1970

PERCENTAGE OF FAMILIES EARNING LESS THAN $4,000

PERCENTAGE OF POPULATION AGE 70+

Since these measurements are used as predictors of case mix composition in the exogenous approach outlined above, they will not be repeated in that approach.

V. Evaluation

It would be desirable if the grouping system, once constructed, could be evaluated as to its "goodness" and/or compared to some other grouping system. Obviously, the criteria by which the system is evaluated depend upon the objectives of the system. That is, if the objective of the grouping system is to find the most statistically pleasing groups (i.e., those groups that minimize intra-group variation of the control parameter--in this case, cost per unit), then evaluation of the proposed system amounts to testing the intra-group variation and comparing it to the variation arising from some other system.[30] However, if the objective is to group hospitals according to variables that in a smoothly functioning market would yield intra-group homogeneity of cost per unit, the statistical evaluation of the system becomes almost impossible. That is, the observation that intra-group variation is smaller in some other grouping scheme cannot be taken as evidence that it is a "better" system given this objective. If the market were functioning smoothly, no control program would be necessary.

However, since the evidence indicates that there are imperfections in the industry, there can no longer be the expectation that hospitals grouped according to the criteria developed in this paper will exhibit, at the start of the program, similar levels of cost per unit. The theoretical implication is that they should, and the purpose of the control program is to see that over time they do. Thus, the grouping system developed here must be evaluated on the basis of its conceptual strength and the translation of the conceptual framework into an implementable program.

---

[30] The best system by this definition includes cost per case unit as a grouping variable.

CHAPTER THREE

CLASSIFICATION METHODOLOGY

I. Introduction

A system of prospective reimbursement discussed in the previous chapter consists of three major elements: (1) selecting appropriate payment units (e.g., per diem costs, per case costs, etc.) and reimbursable costs, (2) classifying hospitals into groups homogenous by external factors, and (3) establishing reimbursement formulas for each identified hospital group on the basis of determined payment units. The previous chapter addressed the first aspect and identified the relevant economic factors which can be used to classify hospitals (See Table 2.3); this chapter will describe the statistical methodology, generically known as cluster analysis, which will arrange hospitals into groups such that hospitals in the same group are more alike with respect to these factors than hospitals in different groups.

The classification methodology consists of the four parts represented in Figure 3.1. The initial part requires determining which constraints, if any, will be imposed a priori on the classification process; e.g., is there a minimum or maximum number of groups required, are some hospitals not allowed in the same group, is there a limit imposed on the number of singly grouped hospitals (isolates), etc.?[31] Constraints which may be imposed in this part of the process might be a function of the type of reimbursement system adopted. For example, if rate setting formulas are based on the performance of hospitals in a group, it may be imperative that each group contain a minimum number of hospitals in order to have a significant sample. In addition, the number of isolates might be constrained, for example, in order to limit the number of hospitals which would have to be treated on an individual basis. Other constraints may be dictated by practical considerations of administering the reimbursement system.

The second part entails calculation of the similarity measures; i.e., the calculation of precise quantities measuring homogeneity between all pairs of hospitals based on the relevant economic variables previously identified. The third part of this study utilizes the similarity measures to determine a hierarchy of groups (called a dendrogram). This hierarchy displays the resultant progress of the cluster analysis algorithm as it proceeds in combining individual hospitals to (ultimately) one group of all sampled hospitals.

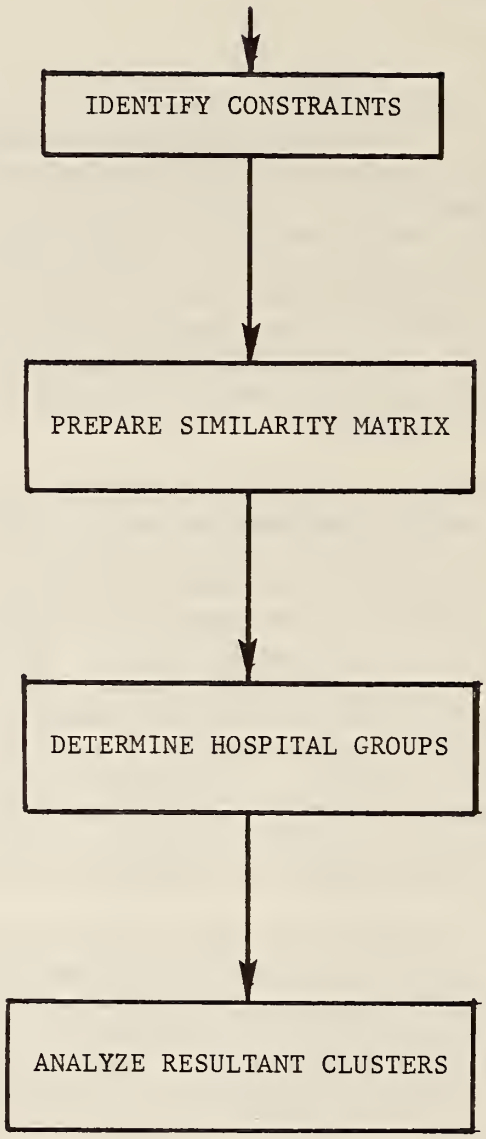
The final step analyzes this hierarchy of hospitals to determine the best hospital partition, the tradeoffs between the number of groups and total homogeneity, and statistical validation of these groups using both parametric and non-parametric procedures. Again, the type of reimbursement system adopted may affect this process of analyzing the hierarchy of hospitals. For example, if

---

[31] For this study, no constraints were imposed other than restricting our examination to short-term general hospitals--a function of the sample constituting our data base.

FIGURE 3.1

Clustering Methodology



the economic framework requires that hospitals in different groups should have different cost structures, then hospitals can be combined to the point where the remaining groups all have significantly different cost structures.

A. Problem Definition

To examine the concept of a cluster hierarchy or dendrogram in more detail and to define the clustering problem more precisely, a formal statement of the problem can be given. Assume there exists a set of n hospitals $H = [H_1, \ldots, H_n]$, where each hospital H_i is described by a (p x 1) vector of variables (listed in Table 2.3), $Y_i = [y_{i1}, y_{i2}, \ldots, y_{ip}]$. For example, y_{i1} might represent input factor prices for the i^{th} hospital, y_{i2} might represent the degree of unionization, etc. Since no unique measurement is available for each variable, several measures are used as surrogates (indicated in Table 2.4); thus, each variable y_{ij} for the i^{th} hospital is represented by a vector of measures or characteristics $[x_{ij1}, x_{ij2}, \ldots, x_{ijq_j}]$, where q_j is the number of measures used for the j^{th} variable.

Given these vectors of measures and weights corresponding to the variables, w_j, we then wish to find a set $G = \{G_1, G_2, \ldots, G_k\}$ of k groups or clusters such that hospitals in the same group are more "alike" than hospitals in different groups with respect to the weighted measures, where the set G partitions H; i.e.,

$$G_q \subseteq H \quad \text{for } q \in Q = \{1, 2, \ldots, k\},$$

$$\bigcup_{q \in Q} G_q = H$$

$$\text{and} \quad G_q \cap G_s = \phi \quad \text{for all } q \neq s \in Q$$

(that is, all hospitals will be placed in a group and no hospital will be in more than one group).

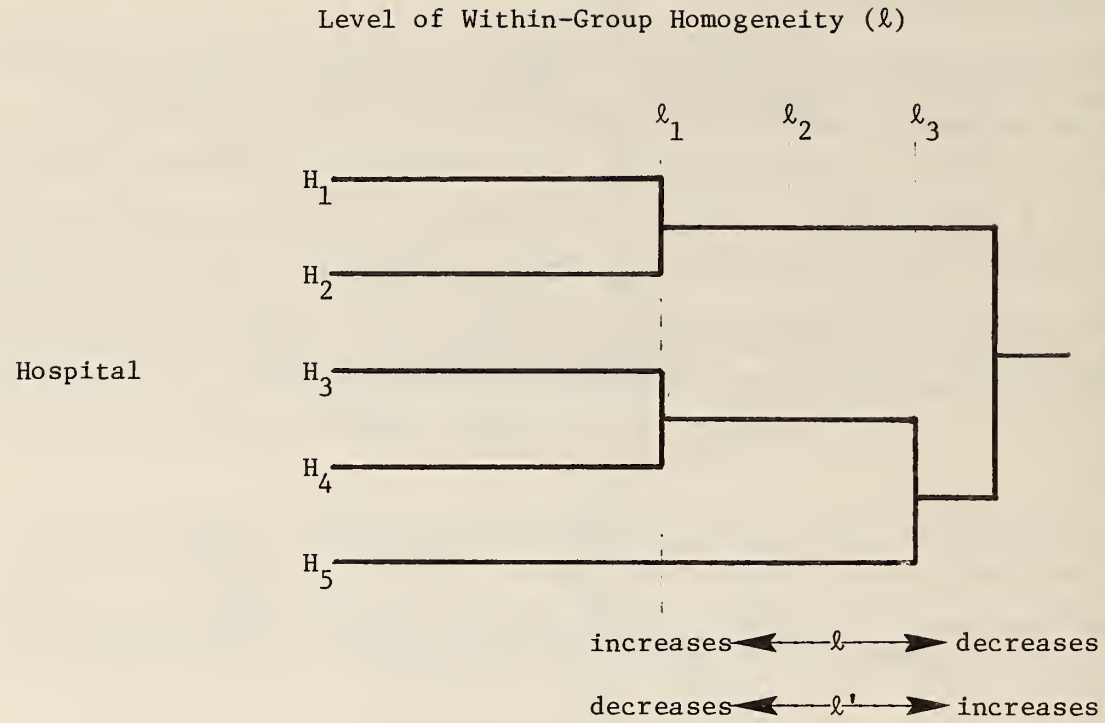
To examine the relationship between the level of homogeneity and the number of groups (k), it is helpful to examine the results of a typical hierarchical clustering scheme.[32] If the number of groups (k) is not specified beforehand, a hierarchical scheme will generate a pattern of cluster combinations often called a dendrogram. While specific measures are discussed below, it follows from the definition of clustering itself that any monotonically increasing measure of homogeneity within clusters (ℓ) will decrease as more hospitals are combined and a similarly defined measure of heterogeneity within clusters (which we will denote ℓ') will increase as combinations take place. An example of such a dendrogram representing five hospitals is presented in Figure 3.2.

---

[32] For the purpose of this discussion, we will represent the level of within group homogeneity as "ℓ".

FIGURE 3.2

Dendrogram Definition



Thus, there exists a relationship between the level of within-group homogeneity (ℓ) and the number of groups (k) for any dendrogram generated by a hierarchical algorithm. If k is predetermined,

$$\ell(k) = f(G)$$

and if ℓ is predetermined,

$$k(\ell) = g(G).$$

For example, given a level of homogeneity ℓ_1, a vertical line drawn through the dendrogram at that point indicates three groups; conversely, assuming three groups a priori indicates a range of homogeneity from ℓ_1 to ℓ_3. Determining these functions, $f(\cdot)$ and $g(\cdot)$, defines the task of the clustering methodology.

II. Measurement Selection, Weighting, and Similarity Measure Computation

As previously indicated, only those variables which cause "legitimate" cost differences should be used for determining homogeneous hospital groups in order for a reimbursement system, based on those groups, to be effective. Thus, having identified the relevant group of exogenously determined variables, the selection, weighting, and use of measurements describing those variables will be discussed here.[33]

Determining the relative importance of grouping variables, as represented by their respective weights, w_c, is a tenuous task at best. In general, weights may be derived from two sources: (1) explicit weights added subjectively by the analyst and (2) implicit weights determined by differences in measurement scales, multicollinearities, etc. Most authors (Anderberg, 1973; Sokal, 1974) argue that all implicit weights should be removed before the clustering process begins; therefore, preclustering calculations will attempt to equalize the relative importance of all selected characteristics to unit weight. Subsequently subjective weighting schemes can be tested for determining the ex post facto sensitivity of resultant cluster definitions.

A. Hospital Similarity Measures Defined

The clustering process is most often initialized by calculating a matrix of similarity measures between all pairs of hospitals being clustered, where the similarity measure between any two hospitals represents a composite score based on the selected characteristics or measurements describing each hospital. While an almost unlimited variety of similarity measures has been suggested (for a complete discussion, see Sokal and Sneath, 1975), many similarity measures can be dismissed on theoretical grounds. For example, a large number of measures (known as coefficients of association) have been developed for dealing with strictly dichotomous data. Other measures, based on the product-moment correlation coefficient, present interpretation problems and are rarely used for comparing objects.[34]

---

[33] As commonly used, these measurements or characteristics themselves may be considered to be variables. In order to avoid confusion, however, we will continue to use the term "variable" to describe the factors indicated by the main headings of Table 2.3, and the interchangeable terms, "measurement" or "characteristic" to indicate those specific empirical quantities which represent the variable in any particular instance.

[34] The exact meaning of the correlation coefficient between two hospitals would be difficult to ascertain; the correlation coefficient implies that a percentage of variation in one hospital's characteristics can be explained by variation in another hospital's characteristics. If we are comparing two hospitals's sizes, for example, it would assume that one hospital's size is some function of the other's--a tenuous assumption at best. Furthermore, an absolute comparison between variables cannot be made; a correlation measure gives a relative comparison only. If two vectors are parallel the correlation coefficient will be $|1|$ --even though the distance (and hence dissimilarity) between the vectors (and hospitals) may be quite large. Furthermore, vectors do not need to be parallel for the correlation coefficient to be equal to unity; as long as some linear relationship exists between the two, the correlation coefficient will be equal to $|1|$.

Given the continuous nature of the data set here, each hospital can be described by a vector of m (where $m = \sum_{j=1}^{p} q_j$) identified characteristics and represented by a single point in m-space, which corresponds to the respective values for the m characteristics. Then, a similarity measure between any two hospitals, say H_r and H_s, can be defined as the distance D(r,s) between the r^{th} and s^{th} points representing the respective hospitals. Distance functions D(r,s), however, are not uniquely defined; however, if they meet the following three criteria,

$$D(r,s) = 0 \quad \text{if and only if } H_r = H_s \tag{3.1}$$

$$D(r,s) = D(s,r) \tag{3.2}$$

and

$$D(r,s) \leq D(r,t) + D(t,s) \tag{3.3}$$

then the distance function is said to be a metric.

The best known and most widely studied metrics are the Minkowski metrics; for hospitals H_r and H_s, each described by an (m x 1) vector of characteristics,

$[x_{r11}, \ldots, x_{r1q_1}, x_{r21}, \ldots, x_{r2q_2}, \ldots, x_{rp1}, \ldots, x_{rpq_p}]$ and

$[x_{s11}, \ldots, x_{s1q_1}, x_{s21}, \ldots, s_{x2q_2}, \ldots, x_{sp1}, \ldots, x_{spq_p}]$, respectively; the

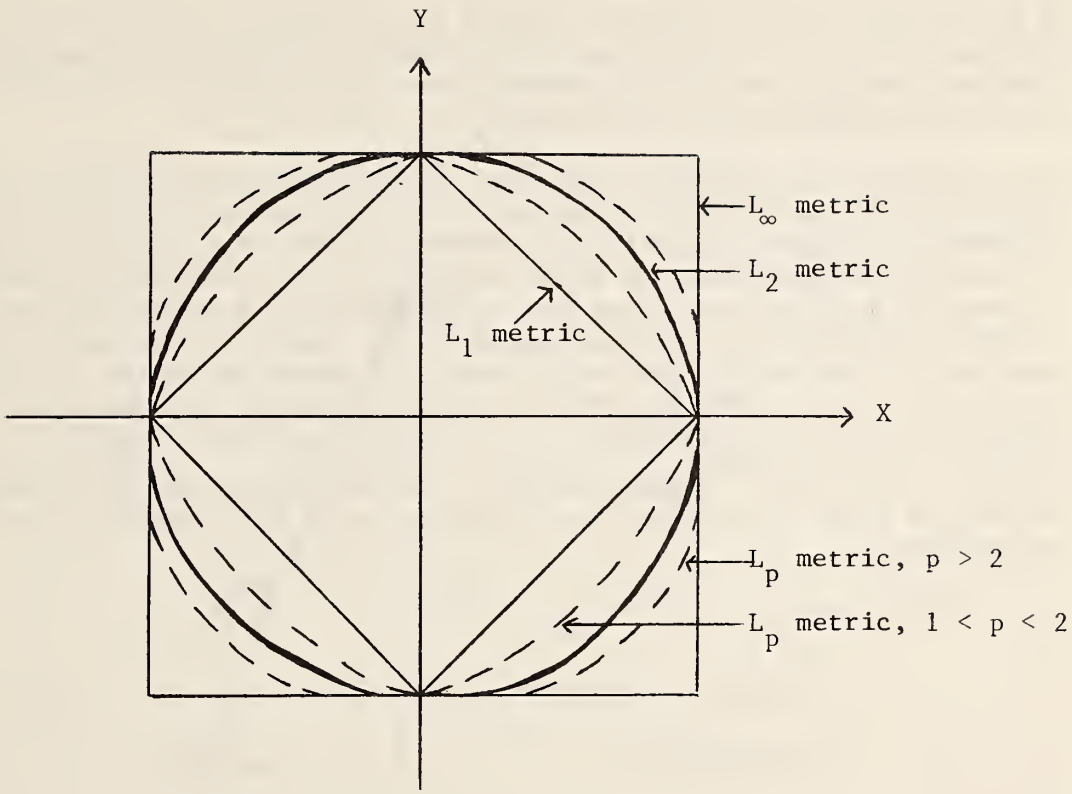
Minkowski metric is defined as follows:

$$d(r,s) = \left[\sum_{j=1}^{p} \sum_{c=1}^{q_j} |x_{rjc} - x_{sjc}|^P \right]^{1/P} \quad \text{for } P \geq 1.$$

Obviously, there are any number of Minkowski metrics for values of $P \geq 1$. Differences between the metrics can be displayed by a graph showing unit distance (the so called "unit ball") for any two points x and y; that is, all points for which the distance from the origin is 1 (Figure 3.3).

FIGURE 3.3

Unit Ball



The three metrics indicated by the solid lines (the L_1, L_2, and L_∞ matrices) have been subjected to the most study and examination; these are described below.

1. The L_1, or "taxicab" or "Manhattan" metric, can be stated for P = 1 as follows;

$$D(r,s) = \sum_{j=1}^{p} \sum_{c=1}^{q_j} |x_{rjc} - x_{sjc}| \ .$$

2. A second metric, letting P = 2, is probably the most familiar as the measure of Euclidean Distance,

$$D(r,s) = \left[\sum_{j=1}^{p} \sum_{c=1}^{q_j} (x_{rjc} - s_{sjc})^2 \right]^{1/2} \ . \qquad (3.4)$$

3. A third metric (letting $P \to \infty$) is sometimes referred to as the Chebychev metric or uniform metric and can be stated as,

$$D(r,s) = \max_{j,c} | x_{rjc} - x_{sjc} |.$$

A variation of the Euclidean distance is calculated by squaring (3.4); this squared distance is widely used, intuitively justified, and analogous to the Mahalanobis distance (discussed later) and, therefore, will form the basic similarity measure used in this study.

To illustrate this distance measure, assume that six hospitals are described by two variables--median county family income and the number of quality enhancing services. In this case, hospitals can be represented by six points in two dimensional space as shown in Figure 3.4. Assuming that each variable is equally weighted and using the squared Euclidean distance (i.e., P = 2) to measure similarity, it is apparent that the hospitals represented by points near each other are similar with respect to these two characteristics, and hence, will have a smaller distance than dissimilar hospitals (for example, hospitals H_2 and H_3 in Figure 3.4 are more alike than hospitals H_2 and H_5 since $D^2(2,3) < D^2(2,5)$). In this example, the squared Euclidean distance between hospitals H_4 and H_6 is easily calculated, using unit weights on each variable, as follows:

$$\begin{aligned} D^2(4,6) &= (I_4 - I_6)^2 + (Q_4 - Q_6)^2 \\ &= (17{,}000 - 5{,}000)^2 + (5-3)^2 \\ &= 144{,}000{,}004. \end{aligned}$$

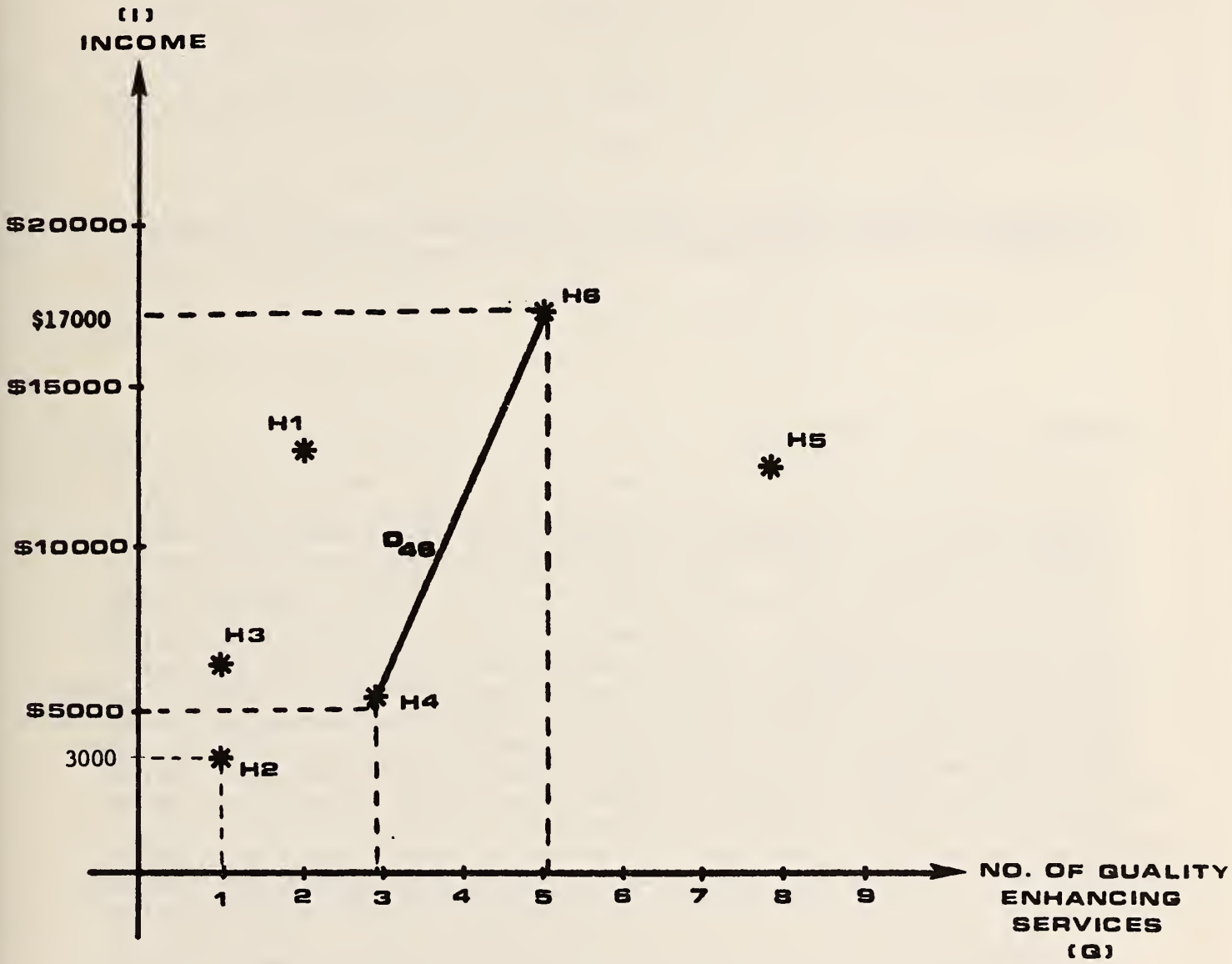
B. Distance Measures Refined

While distance measures are computationally and intuitively straightforward, a number of problems may arise if distance measures are computed directly from the data. ONe apparent problem is caused by differences in measurement scales. In the example in Figure 3.4, one characteristic is measured in terms of number of services provided while the other is measured in dollars--hardly comparable measurement scales. The consequence, therefore, is that the income variable dominates the calculation of $D^2(4,6)$, even though the two variables are theoretically weighted equally.

To accommodate scale differences, a number of possibilities exists. The most commonly utilized approach is to standardize the raw data (that is, subtract the characteristic mean and divide by the characteristic standard deviation) before computing distances between hospitals. Calculating the mean μ_{jc} and

FIGURE 3.4

Hospital Representation



unbiased standard deviation $\hat{s}^2_{jc}$ for the c^{th} measure of variable j,

$$\mu_{jc} = \sum_{i=1}^{n} \frac{x_{ijc}}{n} ;$$

$$\hat{s}_{jc} = \frac{\sum_{i=1}^{n} (x_{ijc} - \mu_{jc})^2}{n-1}$$

the squared Euclidean distance between the r^{th} and s^{th} hospitals can be defined in terms of the standardized measure z_{ijc} as follows

$$D^2(r,s) = \sum_{j=1}^{p} \sum_{c=1}^{q_j} (z_{rjc} - z_{sjc})^2 \qquad (3.5)$$

where $$z_{rjc} = \frac{x_{rjc} - \mu_{jc}}{\hat{s}_{jc}} . \qquad (3.6)$$

Another significant although less obvious problem occurs when two or more characteristics which describe a variable are correlated; metric distances assume an orthogonal space. When spaces are not orthogonal, distances calculated do not follow exactly from equation (3.5). In the example illustrated in Figure 3.4, assume that a third measure, say retail wage rate, is included as an additional measure of the income variable. In this case, it is most likely that the retail wage rate and median family income are highly correlated, and one underlying wage/income factor would explain the variance in both measures. Then, since two wage/income measures are being used, the distances and similarities between hospitals would be unduly weighted in that direction. Thus, it is important to detect any multicollinearities between the measures in order to remove any implicit weighting in the data.

The problem or correlated characteristics was resolved by principal components analysis, a technique which extracts those key factors which are independent or orthogonal of each other[35] (in the above example, the two income measures are combined into a single factor). If there were only one measurement for each theoretically defined variable and we knew a priori the empirical production function and demand curve facing the hospital, there would be no need to extract orthogonal factors; economic theory would dictate the selection and weighting of each variable and any multicollinearities would merely be interesting artifacts of the data. However, given the imperfect data set used in this study and the multiplicity of measures for several variables, it became imperative to examine the measures using principal components analysis in order to test the hypothesized relationship between the measures and their respective variables. Given that relatively well defined factors were found in this study, these factors were used to represent the variables and compute the distance scores.

---

[35] For a more complete description of principal components analysis, see Harmon (1967).

An alternative procedure is based upon the use of the generalized Mahalanobis distance $D^2(i,j)$ (Mahalanobis, 1936), which is defined between hospitals r and s as follows:

$$D^2(r,s) = \Delta_x^T [S^2]^{-1} \Delta_x$$

where S^2 is the variance-covariance matrix and Δ_x is an (m x 1) vector of characteristic differences,

$$\Delta_x = [(x_{rj1} - x_{sj1}), \ldots, (x_{rjc} - x_{sjc}), \ldots, (x_{rjn} - x_{sjn})]$$

The Mahalanobis distance is appealing in that, given a completely orthogonal space, the covariance matrix S^2 reduces to a diagonal matrix of characteristic variances. In this case,

$$D^2(r,s) = \sum_{j=1}^{p} \sum_{c=1}^{q_j} (z_{rjc} - z_{sjc})^2$$

where z_{rjc} is the standardized characteristic defined in (3.6). The Mahalanobis distance then reduces to the standardized squared Euclidean distance defined in (3.5) which was the similarity measure of choice in this study (note that if the characteristics are standardized before computing D^2, the covariance matrix S^2 reduces to the identify matrix and $D^2 = \Delta_x^T \Delta_x$). In general spaces when measurements may be correlated, multicollinearity effects are deleted by the term $[S^2]^{-1}$ (in general, the greater the correlation between two characteristics, the smaller the inverse covariance weighting). Thus, $D^2(r,s)$ is equivalent to standardizing the data, finding orthogonal factor scores from a principal components analysis, and computing squared Euclidean distances. Theoretically, the use of the Mahalanobis distance is advantageous due to its ability to use one hundred percent of the variance; invariably, principal components analysis results in some information reduction. Phillip and Iyer (1974) used both principal components analysis and the Mahalanobis distance; after utilizing principal components analysis to reduce the original set of characteristics, the Mahalanobis distance was subsequently computed in place of the metric distances previously suggested.

In this study, both approaches were tried. However, when the Mahalanobis distance was computed, it was found that the inclusion of all measurement variables resulted in significant instabilities in the data. The problem is caused by the fact that when covariance measures are linearly dependent, the variance-covariance matrix in (3.7) has less than full rank and therefore cannot be inverted. Since subsequent factor analysis showed strong linear depandencies in the variance-covariance matrix, $[S^2]^{-1}$ was calculated but resulted in meaningless values. Therefore, the approach used by Phillip and Iyer (1974) for direct distance calculations from factor scores would have had to be adopted. Similarity measures, however, were calculated from both approaches (i.e., Mahalanobis distance and the squared Euclidean distance calculated from factor scores) in order to verify the distances determined (the two approaches resulted in values which were, in fact, exceedingly close).

C. Similarity Measures - Summary

Measures were standardized and factor scores computed in order to remove the effects of scale differences and multicollinearities among measures, respectively. Since several measures were used to represent most key classification variables, failure to compute orthogonal distances would have given greater weight to the correlated measures.

Once orthogonal factors are found (and implicit weights removed), explicit weights could be added to the factors to take account of the fact they key variables, now represented by the factors, do not have equal impact on hospital cost structure. Two sets of weights were used for each approach. The first set used unit weights to reflect equal weighting on all variables; the second set of weights was determined from regression analysis, using cost per case as the dependent variable. (The precise calculation of these latter weights is described in the following chapter.)

In all cluster analyses performed in this study, similarity measures between hospital pairs were computed using the following steps:

1. Find the mean and standard deviation for each characteristic and compute standardized measures Z_{ijc} from (3.6).

2. Factor analyze the standardized characteristics; reject any factors with eigenvalues less than 1.0.[36]

3. Using the matrix of reduced factor score coefficients and exogenously determined factor weights, calculate the weighted factor scores for each hospital.

4. Using the weighted factor scores, find the squared Euclidean distance between all pairs of hospitals from (3.5).

These steps result in a square matrix of order n containing distance measures representing similarities between all pairs of hospitals in orthogonal space. Since the matrix is symmetrical by (3.2) and the diagonal elements are zero by (3.1), only the $\frac{n(n-1)}{2}$ elements in the upper triangular part of the matrix need to be calculated and retained. An example of such a similarity matrix, which will be used as an illustration in the following sections, is shown in Figure 3.5.

---

[36] Eigenvalues may be loosely interpreted as a measure of explained variance; an eigenvalue cutoff point of 1.0 (percent) was arbitrarily selected to agree with the default value specified by the SPSS (Statistical Package for Social Sciences) program.

FIGURE 3.5

Similarity Matrix

	H_1	H_2	H_3	H_4	H_5	H_6
H_1	--	16.0	15.0	15.0	35.0	32.0
H_2		--	12.5	13.0	38.0	40.0
H_3			--	13.5	36.0	32.0
H_4				--	35.0	34.0
H_5					--	18.0
H_6						--

III. Cluster Dendrogram Determination

Given the hospital representations and the matrix of similarity measures described in the previous section, the next step of the clustering methodology is hospital group determination. Even knowing these representations, however, the problem of detecting homogeneous groups may still be a most problematic one. For example, Figure 3.6 illustrates these cases which may occur: (1) a number of distinct groups exist which can be delineated by linear functions, (2) a number of distinct groups exist which cannot be delineated by linear functions, and (3) no distinct groups appear evident other than the single isolated hospital. While most classification problems encountered in the real world fall into the third case, even problems in the first two cases remain exceedingly difficult to resolve. To illustrate this difficulty, examine the first case in Figure 3.4, where the number of hospitals (6) and

apparent groups (2) is limited. In this limited case, searching all possible partitions using, say, linear discriminant analysis, would entail examining 31 possible partitions (enumerated in Table 3.1).[37] Not knowing the value of k (the number of groups) beforehand, one would have to examine the sum of a series of Stirling numbers of the second kind; in this example, k would vary from 1 to 6 and a total of 203 possible partitions would have to be examined. For this reason of combinatorial complexity, only heuristic algorithms will be used in order to maintain computational feasibility.[38]

Heuristic clustering strategies can be subdivided into hierarchical methods, iterative methods, and ad hoc methods as represented in Figure 3.7. Hierarchical methods either begin with all hospitals in individual groups and subsequently combine these groups in some fashion (agglomerative clustering), or initially place all hospitals in one group and subsequently split the initial and following groups in such a way as to satisfy some criterion at each stage (divisive clustering). Agglomerative clustering begins with n groups and ends with one; divisive clustering starts with one group and ends with n. Agglomerative methods (represented in Figure 3.2) can be further subdivided on the basis of clustering criteria; linkage methods examine the total of some within group measure. Iterative methods, on the other hand, normally require assuming a priori a value of k (the number of groups) and begin with some arbitrary partition of hospitals into k groups and proceed to rearrange the hospitals in order to decrease the level of homogeneity among all groups.

Given that divisive methods have some theoretical disadvantages which make them less desirable than agglomerative methods (Gower, 1967), the use of such

---

[37] In general, for n hospitals and k groups, the number of possible partitions is equal to S_n^k, a Stirling number of the second kind, where

$$S_n^k = \frac{1}{k!} \sum_{p=0}^{k} (-1)^{k-p} \binom{k}{p} p^n$$

and $\binom{k}{p}$ is the number of combinations of k objects taken p at a time, equal to $\frac{k!}{p!\,(k-p)!}$.

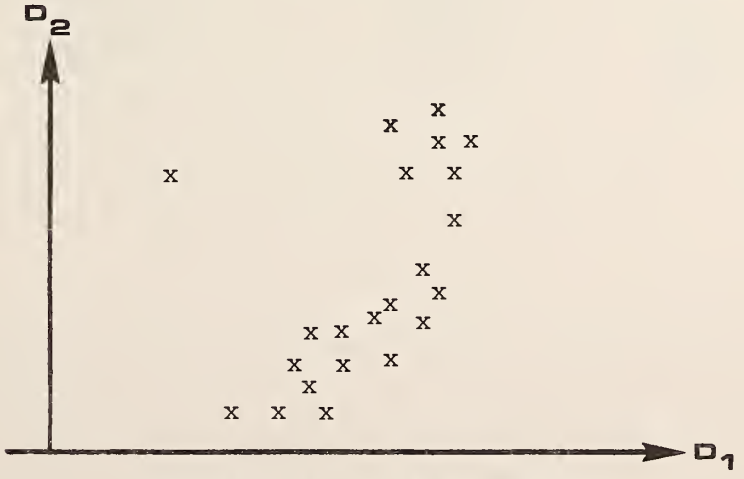
[38] While a number of optimization approaches to the clustering problem have been proposed, none has been demonstrated to be computationally feasible for anything other than trivial problems. For a description of several suggested optimization algorithms, see Klastorin (1973).

FIGURE 3.6

Classification Illustrations

Three groups - distinct - linear

Two groups - distinct - nonlinear



Nondistinct groups - isolate

TABLE 3.1

Possible Partitions

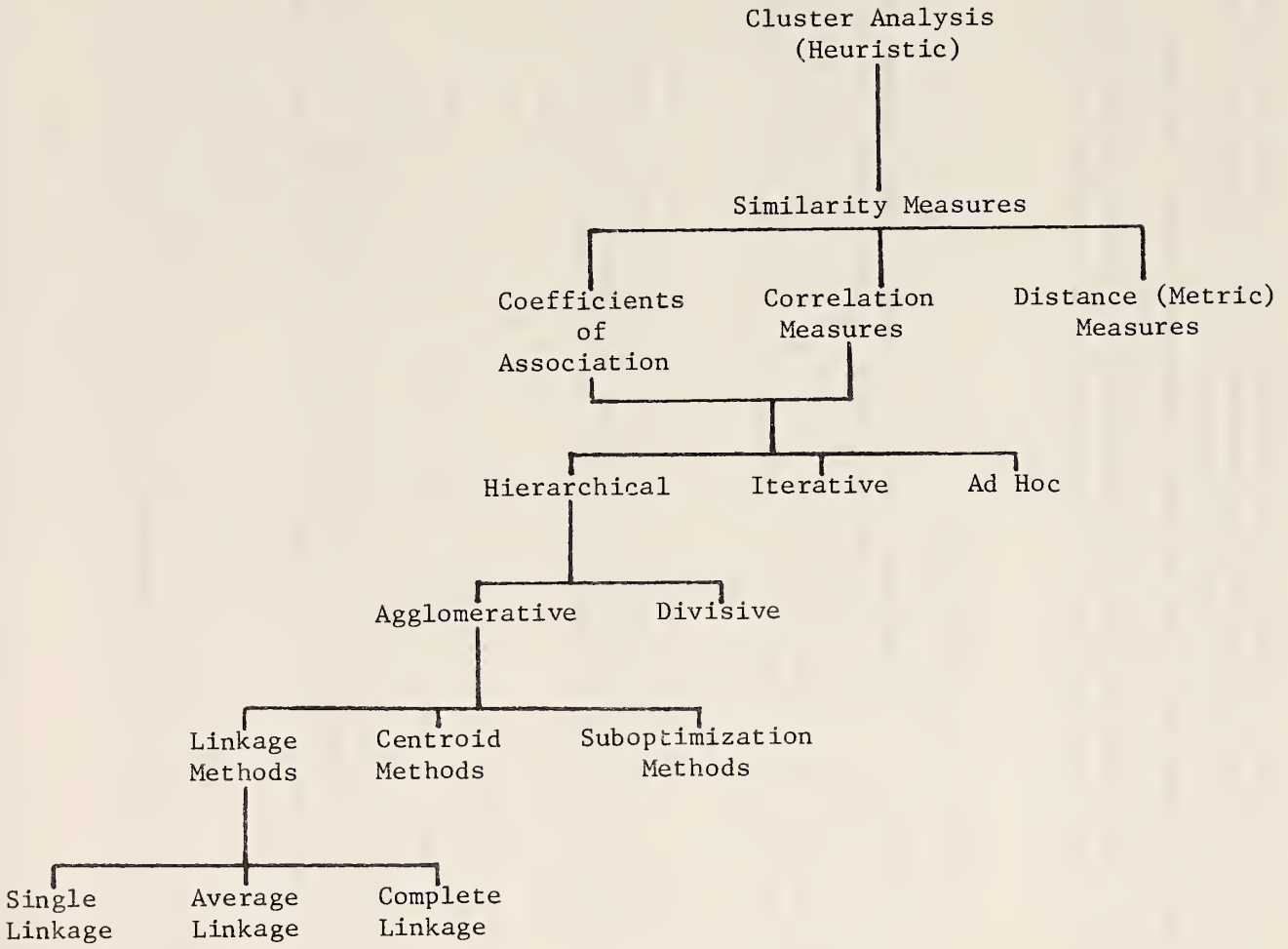
(N = 6)

TWO GROUPS

	Group 1	Group 2
1.	(1)	(2,3,4,5,6)
2.	(2)	(1,3,4,5,6)
3.	(3)	(1,2,4,5,6)
4.	(4)	(1,2,3,5,6)
5.	(5)	(1,2,3,4,6)
6.	(6)	(1,2,3,4,5)
7.	(1,2)	(3,4,5,6)
8.	(1,3)	(2,4,5,6)
9.	(1,4)	(2,3,5,6)
10.	(1,5)	(2,3,4,6)
11.	(1,6)	(2,3,4,5)
12.	(2,3)	(1,4,5,6)
13.	(2,4)	(1,3,5,6)
14.	(2,5)	(1,3,4,6)
15.	(2,6)	(1,3,4,5)
16.	(3,4)	(1,2,5,6)
17.	(3,5)	(1,2,4,6)
18.	(3,6)	(1,2,4,5)
19.	(4,5)	(1,2,3,6)
20.	(4,6)	(1,2,3,5)
21.	(5,6)	(1,2,3,4)
22.	(1,2,3)	(4,5,6)
23.	(1,2,4)	(3,5,6)
24.	(1,2,5)	(3,4,6)
25.	(1,2,6)	(3,4,5)
26.	(1,3,4)	(2,5,6)
27.	(1,3,5)	(2,4,6)
28.	(1,3,6)	(2,4,5)
29.	(1,4,5)	(2,3,6)
30.	(1,4,6)	(2,3,5)
31.	(1,5,6)	(2,3,4)

FIGURE 3.7

Cluster Analysis (Heuristic)



algorithms was not considered. On the other hand, iterative algorithms would be relatively undesirable for examining the tradeoffs between the number of groups and total homogeneity. Another disadvantage encountered with iterative methods is imposed by the nature of the iterative algorithms; most[39] are based on simple pairwise exchanges from an initial (and often arbitrary) partition. Recent evidence (Cormack, 1973) indicates that these procedures may frequently terminate at poor solutions.

On balance, heuristic agglomerative algorithms appear most appropriate for use in an examination of grouping patterns. In addition, it would be desirable to use a number of these clustering strategies for the purpose of establishing consistency requirements and validating clusters found. In other words, given the heuristic nature of the algorithms used, additional credibility could be attached to groups simultaneously identified by several algorithms. This concept forms the basis for the development of the composite dendrogram, which will be explored in detail in a later section.

A. Agglomerative Clustering

We now turn to the examination of several algorithms which, based on the similarity matrix S, calculate successively inclusive group structures. (Section I briefly alluded to such agglomerative schemes by displaying a typical resultant dendrogram.) following this discussion, it will be shown how the results from these algorithms can be used to compute a composite dendrogram.

A.1. Linkage Methods

Linkage methods are both the simplest and the most widely used clustering methods. Essentially, there are three categories of linkage methods--single linkage, complete linkage, and average linkage--whose differences will be examined in this section.

All linkage methods begin and proceed in a similar manner; the difference between methods is the cricerion used to merge groups of hospitals at given stages. Given n hospitals and a symmetrical distance matrix $|D(r,s)|$, these algorithms begin by assigning each hospital to a distinct group. Letting ℓ be a measure of homogeneity within groups and representing each of the n groups by $\{H_i\}$, the development a typical agglomerative algorithm can be followed by examining the dendrogram in Figure 3.2. As ℓ is increased from zero (or some small number) groups are combined if and only if their respective distances are less than or equal to the value of ℓ. In the procedures used in this study, ℓ is defined by examining the minimum distance required for a given algorithm to combine all hospitals into one group. This value is then divided into 25 equal intervals and ℓ is incremented 25 times; each increment level (labeled from 1 to 25 on the computer output) is referred to as the cluster or class level.

---

[39] Included, for example, would be such well known algorithms as MacQueen's "K-means" technique (MacQueen, 1967), and ISODATA (Ball and Hall, 1965).

A.1.a. Complete Linkage

The complete linkage clustering criteria dictates that group $\{H_r\}$, for example would be combined with group $\{H_s, H_i\}$ if

$$\max \left|D^2(r,s), D^2(r,t)\right| \leq \ell$$

for any level of homogeneity ℓ. Equation (3.8) indicates that any two groups are combined if and only if distances between all pairs of hospitals in the two groups are less than or equal to ℓ. The complete linkage algorithm results in compact clusters; combining groups only when all links between hospitals are less than the level of homogeneity ℓ is equivalent to minimizing the within group diameter at each stage of the algorithm. In this case, each cluster can be pictured as an m-dimensional sphere with the largest intra-cluster distance as the diameter.

The complete linkage algorithm has been widely used and offers the distinct advantage of being invariant to monotone transformations of the data. This property is especially important in this study where credibility can be established only on the ordinal ranking of some measurements. The complete linkage algorithm can be illustrated by the similarity matrix in Figure 3.5 and the dendrogram in Figure 3.8. In this case, a search of the $\frac{n(n-1)}{2}$ similarity measures indicates that the minimum value is 12.5; thus, ℓ must be increased to 12.5 before any of the hospitals (in this case, H_2 and H_3) are grouped together at cluster level 1. Continuing the search, ℓ might be increased to 13.0 (the next smallest value), but no additional grouping would take place as all pairwise distances (i.e., $D^2(2,3)$, $D^2(3,4)$) are not less than or equal to 13.0 (here, $D^2(3,4) = 13.5$). Thus, for another merger to take place, ℓ must be increased to 13.5 and group $\{H_2, H_3\}$ and $\{H_4\}$ will be merged as indicated at level 2 in the dendrogram. The search continues in this manner; when $\ell = 16.0$, the groups $\{H_2, H_3, H_4,\}$ and $\{H_1\}$ are combined as

$$\min \left|D^2(1,2), D^2(1,3), D^2(1,4)\right| \leq 16.0.$$

When $\ell = 18.0$, groups $\{H_5\}$ and$\{H_6\}$ are combined (level 5); all hospitals are not combined until $\ell \geq 40.0$ at level 9. (In the algorithms actually used, the process would stop at level 9 when all hospitals are merged into one group.)

A.1.b. Single Linkage

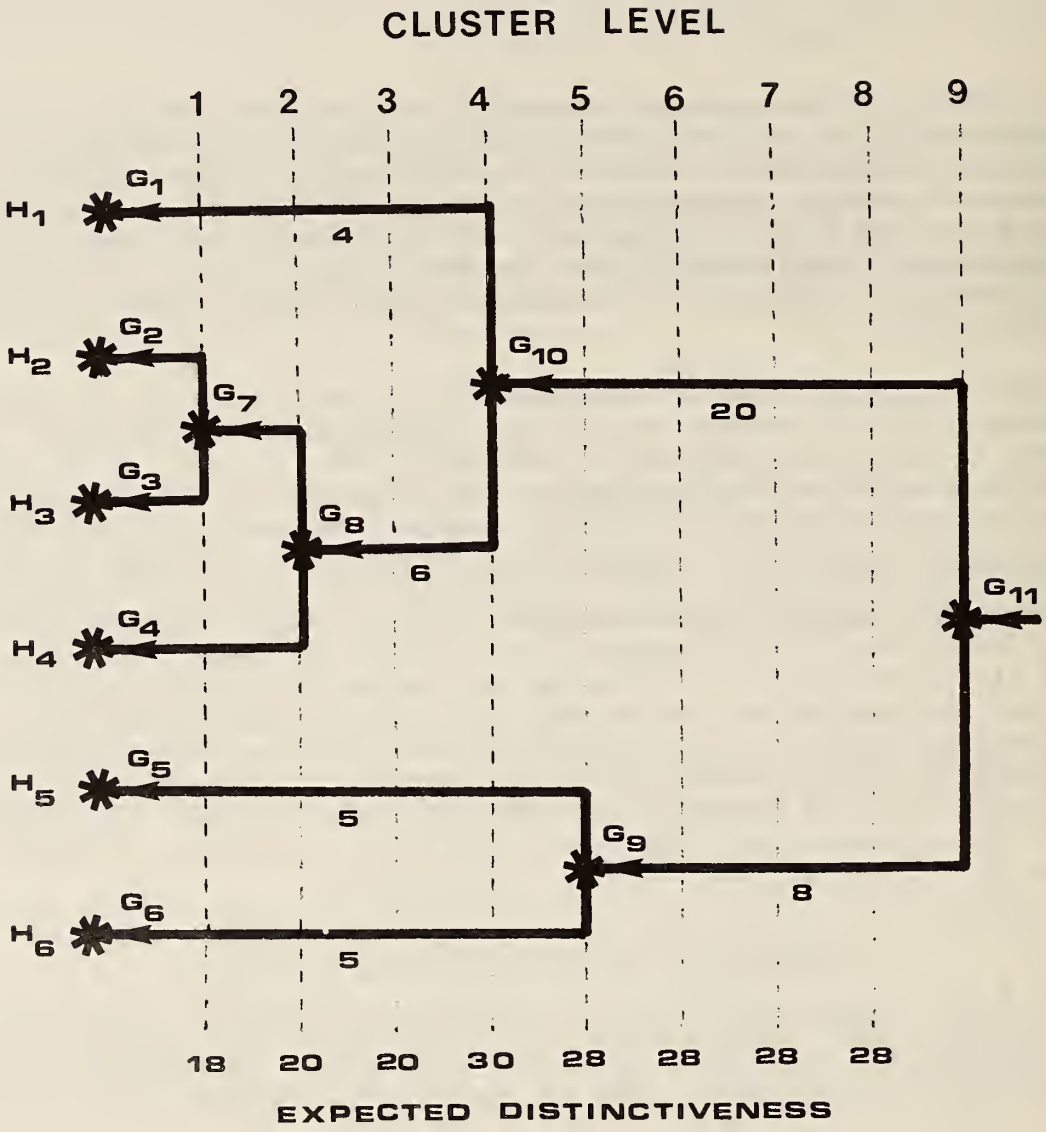
If the clustering criterion (i.e., when groups $\{H_r\}$ and $\{H_s, H_t\}$ are combined) is

$$\min \left|D^2(r,s), D^2(r,t)\right| \leq \ell,$$

the method is known as single linkage. The above equation indicates that two groups will be combined as long as at least one link between hospitals in one group and hospitals in the other group is less than ℓ.

FIGURE 3.8

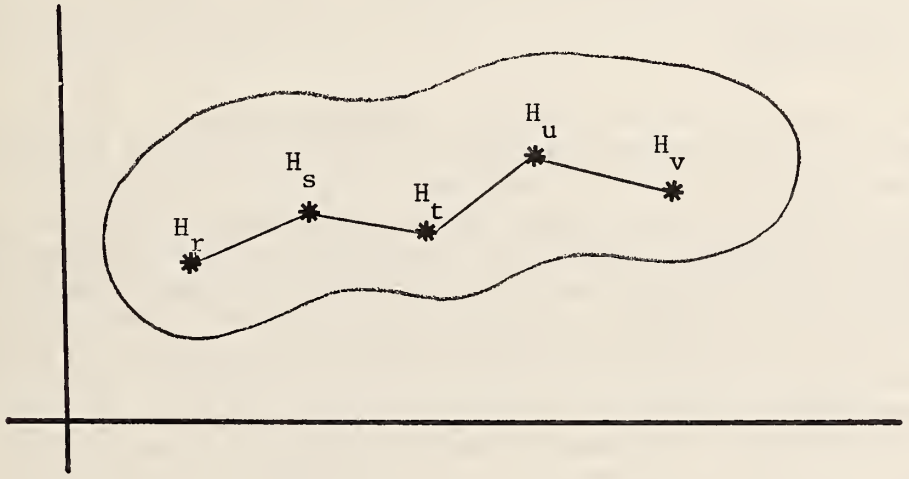
Example Dendrogram for Six Hospitals



Single linkage methods often lead to a result known as chaining. For example, if $D^2(r,s)$, $D^2(s,t)$, $D^2(t,u)$, and $D^2(u,v)$ are less than ℓ, all five hospitals would be combined in one group H_r, H_s, H_t, H_u, H_v even though the distance between hospital H_r and hospital H_v may be as large as 4ℓ (assuming the similarity measure is a metric so that the triangle inequality holds). This cluster is pictured in Figure 3.9.

FIGURE 3.9

Single Linkage Cluster Example



The effect of chaining often leads to a criticism that single linkage methods, while not necessarily resulting in compact clusters, do not yield sufficient information about the structure of the cluster itself (Wishart, 1970). While some controversy does exist with respect to the usefulness of these serpentine clusters (Cormack, 1973), the decision was made to exclude the use of this algorithm in this study. Given the simultaneous use of multiple algorithms and the nature of the hospital data, it appears that a single linkage approach would be inappropriate.

A.1.c. Average Linkage

If two existing clusters are merged when the average distance measure within a new cluster is less than ℓ, the process is called average linkage.

A variation of this criterion has been suggested by Lance and Williams (1966), who proposed to examine only links between the two candidate groups and merge if the average distance is less than ℓ.

Anderberg (1973) reports that the results obtained from these methods are, not surprisingly quite close and, in general, give results not too dissimilar from those obtained by complete linkage methods.

A.2. Centroid Methods

Centroid methods are similar to the concept of average linkage. Initially hospitals $\{H_r\}$ and $\{H_s\}$ are combined if $D^2(r,s)$ is the minimum of all inter-hospital distances; the r^{th} and s^{th} columns (and rows) in the similarity matrix are then replaced by one (average) vector. The process is then repeated on the (n-1) x (n-1) similarity matrix; the two groups with the smallest distance are joined. The distance between clustered groups at each stage provides a measure of homogeneity ℓ; however, centroid methods do not necessarily guarantee that the distance function will monotonically decrease (the cluster centroids may change or float with the merge of each new group). A primary difference between centroid methods and linkage methods is that the former methods describe clusters at any stage by the differences between a single vector of scores (centroids), whereas the latter methods describe clusters by the differences between the elements (i.e., hospitals of the clusters.)

Sokal and Sneath (1973) and Sokal and Michener (1958) were among the first to describe centroid methods based on the arithmetic mean. Replacing several hospitals in two groups by their single joint mean (i.e., centroid) has the effect of weighting each prior group by the number of hospitals in that group.

While Sokal and Michener (1958) argue the plausibility of such weights, Gower (1967) offers a scheme utilizing the median in place of the mean. In the latter approach, if groups G_r and G_s are used to represent the new vector G_t.

A.3. Ad Hoc Methods

There have been few optimization algorithms in hierarchical clustering. Most optimization routines (most notably Ward, 1963; and Ward and Hook, 1963) suboptimize in the sense of minimizing (or maximizing) some criterion at each stage and thereby avoid determining a final, global optimum partition. Mention of optimization desirability was made in 1958 by W. Fisher who, in order to reduce the complexity of the problem restricted himself to analyzing single dimensional vectors (reducing the vector of characteristics to a single measurement x_i for the i^{th} hospital). Assigning the i^{th} hospital to the r^{th} group (G_r), Fisher defined homogeneity, synonymous with cluster compactness, for the p^{th} group as:

$$\xi_r = \sum_{k \varepsilon G_r} (x_i - \tilde{G}_r)^2$$

(where $\tilde{G}_r$ is the mean value for group G_r). The measure of homogeneity expressed above is the sum of the square Euclidean distances for all units in the r^{th} group from the mean G_r, more often simply designated as the "within group sums of squares." From ξ_r, several aggregate measures can be derived, the most obvious being the total within group sums of squares.

$$Z = \sum_{r=1}^{k} \xi_r$$

and, as a second alternative, the average total sum of squares, Z/k.

An algorithm proposed by Ward (1963) generalized Fisher's problem to m dimensions. While Ward's algorithm accommodates any functional, it is best known by his example, minimizing the total sums of squared deviations about the group mean.

Ward's measure of homogeneity ($Z = \sum_{r=1}^{k} \xi_r$) differs from Fisher's by varying the value of k and thereby constituting a hierarchical algorithm (Fisher assumed a fixed k). Ward's algorithm, using Z as a stage-wise criterion, can be expressed as:

$$\text{find min } Z = \sum_{r=1}^{k} \xi_r \quad \text{for all } k = n, n-1, \ldots, 1.$$

Determining a global minimum for Z would require complete enumeration and calculation of all possible values of Z at each stage. Ward's algorithm, however, moves from the k^{th} stage (i.e., k clusters exist) to the $k\text{-}1^{st}$ stage by only examining the $\frac{k(k-1)}{2}$ possible pairwise combinations at each stage reducing the number of clusters at each stage by one. While this procedure results in minimal Z at each stage, it does not guarantee a global minimum. Ward's method, demonstrated by the error sum of squares criterion, may be calculated using various objective functions.

B. Composite Dendrogram Calculation

Most studies using cluster analysis proceed by selecting one of the cluster heuristics described in the previous section--the algorithm selection often based on computer code and/or programming availability (for example, the study by Phillip and Iyer (1972) used the complete linkage agglomerative algorithm). In some cases, results from more than one algorithm are evaluated and, on the basis of some evaluative criteria, one set of results is accepted over another.

Given the combinatorial nature of the basic clustering problem, the heuristic nature of computationally feasible algorithms, and the desire to minimize the probability of incorrect hospital grouping, it was decided here to use a number of algorithms and combine their results in a composite approach. Attempting to select a unique resultant dendrogram based on some evaluation criteria did not prove feasible in this study as it was impossible to detect statistically significant differences between results in most cases. In this instance, a statistically superior approach would be provided if the composite dendrogram were computed such that the probability of misgrouping hospitals was minimized. While the likelihood of grouping two nonhomogeneous hospitals together would be reduced by this conservative approach, it might, on the other hand, result in an increased number of groups. The calculation of this composite dendrogram and an example illustrating its use are presented in the following sections.

B.1. Cophenetic Correlation Coefficient

One measure which may be used to evaluate resultant dendrograms, first suggested by Sokal and Rohlf (1962),[40] is referred to as the cophenetic (i.e., class level) correlation coefficient. Since the hierarchical programs described in the previous section divide the total measure of homogeneity (ℓ) in the resultant dendrogram into 25 distinct intervals or class levels, the level where any pair of hospitals are joined can easily be determined for any algorithm. Then, given the vector of joining class levels for all pairs of hospitals, the cophenetic correlation coefficient can be defined as the correlation coefficient between the vector of class levels and the vector of similarity measures or distances between all hospital pairs.

---

[40] Measures of this type have been used by Boyce (1969) and Green and Carmone (1970).

For the (six hospital) example in Figure 3.8, there are 15 hospital pairs,[41] which are enumerated in Table 3.2. For each hospital pair, one element in the vector of pairwise joining levels is easily determined from the dendrogram in Figure 3.8; i.e., the minimum value of ℓ(level) from 1 to 9 at which the two hospitals are grouped together. For the same pairwise combination, elements of another vector of distance scores are generated from Figure 3.5. The cophenetic correlation coefficient is defined as the Pearson product moment correlation between the two vectors.

Several problems, however, arise from the use of this approach. First, pairwise joining levels are defined only for discrete ordinal values (in Table 3.2, values from 1 to 9) while the distance score vector is defined by continuous interval quantities--thereby creating potential difficulties when computing the correlation between the two vectors (see Feller, 1968). Thus, by using ordinal values of the joining level, which approximates the exact homogeneity level at which each hospital pair was joined, some information was sacrificed. Lastly, use of the Pearson product moment correlation coefficient assumes that the vector elements are drawn from multivariate normal distributions.

TABLE 3.2

Cophenetic Correlation Coefficient: Absolute Validation

Hospital Pair	Pairwise Joining Levels	Distance Scores	Distance Joining Scores
(1,2)	4	16.0	16.0
(1,3)	4	15.0	16.0
(1,4)	4	15.0	16.0
(1,5)	9	35.0	40.0
(1,6)	9	32.0	40.0
(2,3)	1	12.5	12.5
(2,4)	2	13.0	13.5
(2,5)	9	38.0	40.0
(2,6)	9	40.0	40.0
(3,4)	2	13.5	13.5
(3,5)	9	32.0	40.0
(3,6)	9	32.0	40.0
(4,5)	9	35.0	40.0
(4,6)	9	34.0	40.0
(5,6)	5	18.0	18.0

[41] In general, the number of pairwise combinations is ${}_nC_2 = \binom{n}{2} = \frac{n!}{(n-2)!\,2!}$.

The first two problems can be resolved by using the exact value of ℓ when each hospital pair was joined (in lieu of the joining level) and then comparing the vector of pairwise distance scores with this vector of joining distance scores. While there remains a one to one correspondence between the joining level and joining distance score vectors in Table 3.2, the vectors being correlated now both consist of interval quantities. The last problem (i.e., the normality assumption) is easily resolved by using the nonparametric Spearman correlation coefficient. This measure will hereafter be referred to as the cophenetic correlation coefficient.

B.2. Composite Dendrogram Calculation

Using several heuristic algorithms and the square of their respective cophenetic correlation coefficients as measures of validity, the results can be combined into a single composite dendrogram. Basically, the approach here is to group hospitals together only when a "weighted majority" of the algorithms agree that such hospitals are in fact homogeneous. Such an approach has the effect of minimizing the probability that dissimilar hospitals will be grouped together, while allowing the possibility that additional groups will be created. Six heuristic cluster algorithms were used to form a complete composite dendrogram; these algorithms (described in the previous section) included the following:

1. complete linkage,
2. average linkage between merged groups,
3. centroid method
4. median method of Gower,
5. average linkage within groups,
6. Ward's suboptimization method.

To find the composite dendrogram, let Δ be the set of individual algorithms used and r^2_δ (for $\delta \epsilon \Delta$) equal the squared cophenetic correlation coefficient for each respective algorithm (where $0 \leq r^2 \leq 1$). In order to vary the sensitivity of the composite dendrogram to differences among the separate cluster algorithms, a single coefficient β is defined, where the range of β is defined over the discrete interval of class levels (in our study, from 1 to 25). Basically, equals the number of recognized class levels in each dendrogram; for example, if β equals 1 (the least sensitive position) then all levels are considered together and all hospitals are grouped at level 1. If β equals 25 (the most sensitive position), then each joining class is recognized; if β equals 12.5 then class levels 1 and 2 are equated, class levels 3 and 4 are equated, etc. β is also used to define $n_{ij\delta}$, the number of recognized class levels at which the i^{th} and j^{th} hospitals were joined by algorithm δ. For example, if β equals 25, then $n_{ij\delta}$ simply equals the class level at which the i^{th} and j^{th} hospitals join together by algorithm δ; if β equals 12.5, then $n_{ij\delta}$ equals 1 if hospitals i and j are joined at either class level 1 or 2, etc.

Given r^2_δ, β, and $n_{ij\delta}$, a measure d_{ij} can be found which expresses the "distance" or similarity between all pairs of hospitals based on joining class levels for all hospital pairs from individual dendrograms, weighted by their respective

r_δ^2. Such a measure is defined as follows:

$$d_{ij} = 1 - \frac{\sum_{\delta\varepsilon\Delta} (1 - \frac{n_{ij\delta}}{\beta}) r_\delta^2}{\sum_{\delta\varepsilon\Delta} r^2} \quad . \tag{3.9}$$

The range of d_{ij} varies from $\frac{1}{\beta}$ to (for all $r_\delta^2 = 1$); the sensitivity index of β is easily visualized as the determinant of the range of d_{ij} (which increases with the value of β). Thus, a larger value of β results in a larger range of d_{ij}, more discrimination between hospitals, and more sensitivity in the composite dendrogram. β is a useful concept in that many of the groups formed at the initial class levels are two small and numerous to be of much interest. Thus, setting a smaller value of β (say $\beta \simeq 8$) eliminates examination of these structures and concentrates upon dominant structures or partitions. Note that d_{ij} is indeterminant if all r_δ^2 are equal to zero.

Examining the computation of d_{ij} further, it becomes apparent that

$$d_{ij} = 1 - \frac{\sum_{\delta\varepsilon\Delta} r_\delta^2 - \frac{1}{\beta} \sum_{\delta\varepsilon\Delta} n_{ij\delta}\, r_\delta^2}{\sum_{\delta\varepsilon\Delta} r_\delta^2} = 1 - 1 + \frac{1}{\beta} \frac{\sum_{\delta\varepsilon\Delta} n_{ij\delta}\, r_\delta^2}{\sum_{\delta\varepsilon\Delta} r_\delta^2}$$

$$= \frac{1}{\beta \sum_{\delta\varepsilon\Delta} r_\delta^2} \sum_{\delta\varepsilon\Delta} n_{ij\delta}\, r_\delta^2$$

Since $\frac{1}{\beta \sum_{\delta\varepsilon\Delta} r_\delta^2}$ is constant, the term can be ignored and d_{ij} defined simply as

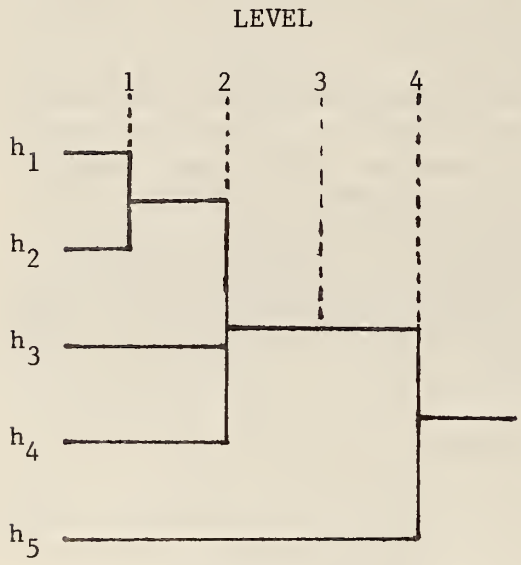
$$d_{ij} = \sum_{\delta\varepsilon\Delta} n_{ij\delta}\, r_\delta^2 \quad . \tag{3.10}$$

In (3.10), d_{ij} now ranges from η to $\eta\beta$ (where η is the number of algorithms) for $r_\delta^2 = 1$; thus, (3.10) is identical to changing the scale of (3.9) by a constant factor of $\eta\beta$.

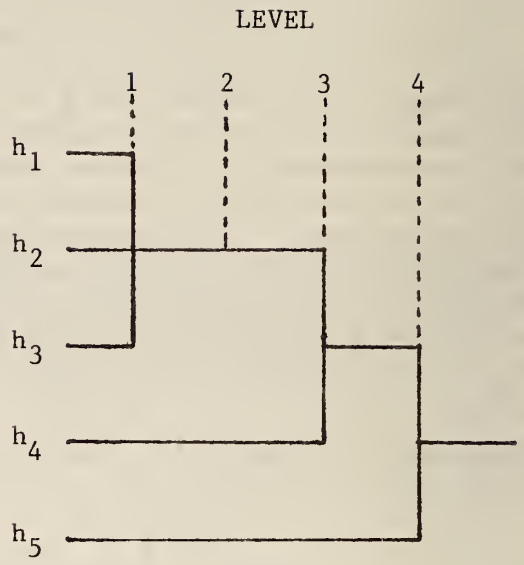
To illustrate the calculation of the composite dendrogram, assume that two algorithms and their respective dendrograms (pictured in Figure 3.10) are to be combined, where $\eta = 2$, $n = 5$ hospitals, and cophenetic correlation coefficients for both dendrograms (r_δ^2) are equal to 1.0. For $\beta = 4$ (most sensitive value) and $\beta = 2$, the matrices of similarity measures are shown in Table 3.3; Figure 3.10 shows the resultant composite dendrograms. As evident from an examination of the resultant dendrograms, a value of $\beta = 4$ results in some discrimination between the grouping of hospitals 1, 2, and 3, while a value of $\beta = 2$ shows no discrimination.

FIGURE 3.10

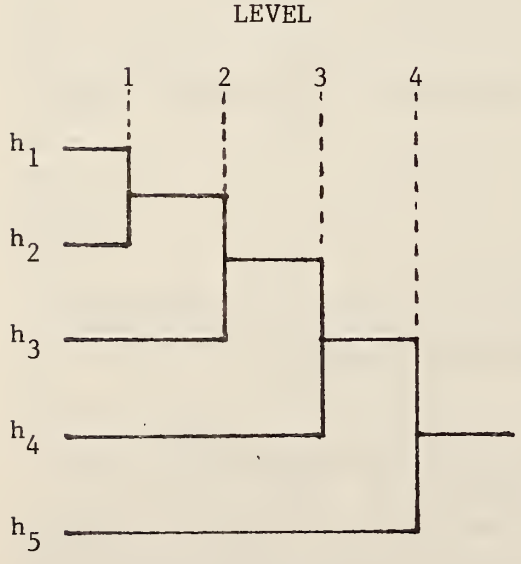
Composite Dendrogram: Example



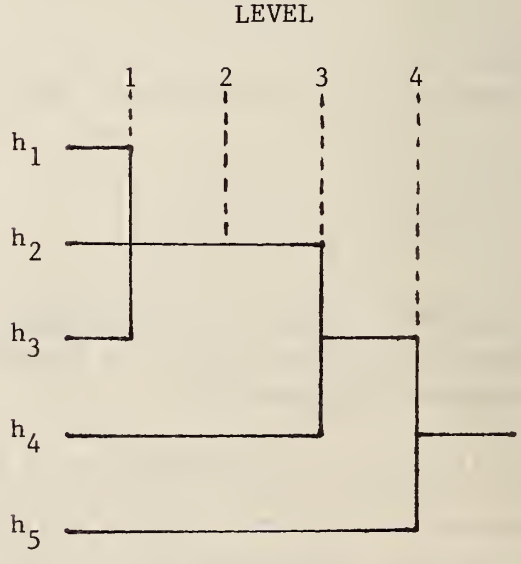
ALGORITHM NO. 1
(R_2^2 = 1.0)



ALGORITHM NO. 2
(R_2^2 = 1.0)



COMPOSITE
(β = 4)



COMPOSITE
(β = 2)

TABLE 3.3

Similarity Measures d_{ij}

(Example: Figure 3.10)

$\beta = 4$

	h_1	h_2	h_3	h_4	h_5
h_1	--	2	3	5	8
h_2		--	3	5	8
h_3			--	5	8
h_4				--	8
h_5					--

$\beta = 2$

	h_1	h_2	h_3	h_4	h_5
h_1	--	2	3	3	4
h_2		--	3	3	4
h_3			--	3	4
h_4				--	4
h_5					--

Note that the procedure for combining dendrograms essentially defines a complete linkage algorithm using d_{ij} as similarity measures. Thus, once d_{ij} values are calculated, the composite dendrogram requires no new computer algorithm.

In practice it was found that value of β equal to 25, 12.5, or 8 resulted in the most useful structures. However, given the development of measures to detect and evaluate dominant partitions within a dendrogram (described in the following section), it became unnecessary to examine various values of β. Thus, due to the desire to minimize the probability of clustering nonhomogeneous hospitals, the value of β was set to 25 for purposes of this study, and only the most sensitive composite dendrograms were used.

IV. Partition Evaluation

Having determined the composite dendrogram, the problem still remains of evaluating the dendrogram and finding which partition offers the best tradeoff between total homogeneity and number of groups. Note that in any dendrogram (Figure 3.8 for example), a partition can be formed by placing a vertical line through any class level. Thus, the initial question is how to objectively evaluate partitions formed (by vertical lines) at each class level, without having to rely upon visual examination.

A. Expected Distinctiveness Defined

Examining a dendrogram carefully, it becomes evident that each horizontal line segment defines a separate group of hospitals (in Figure 3.8, there are 11 different groups labeled G_1 through G_{11}). Each group's distinction or importance, in a sense, is defined by the length of its respective line segment as measured by the number of class levels. For example, group G_2 only exists for one class level while group G_1 exists for four class levels--

implying that G_1 is more distinct than G_2. Given that a probability is a number from 0 to 1 which is the limit of relative frequencies, the probability of any hospital H_i belonging to group G_q can be defined using this concept of distinctiveness.

Therefore, letting

$$s_{iq} = \begin{cases} 1 & \text{if hospital } H_i \text{ belongs to group } G_q \text{ (i.e., if } H_i \varepsilon G_q) \\ 0 & \text{otherwise,} \end{cases}$$

the probability that any hospital H_i belongs to group G_q can be defined as

$$P(s_{iq}) = P(H_i \varepsilon G_q) = \lim_{D \to \infty} \frac{(\text{distance } H_i \text{ belongs to } G_q)}{D}$$
$$= \lim_{D \to \infty} \left(\frac{d(G_q)\ s_{iq}}{D} \right),$$

where

$d(G_q)$ = number of class levels for which group G_q exists

and D = maximum number of class levels.

In the example in Figure 3.8, D=10 and $P(H_1 \varepsilon G_1) = 0.04$, $P(H_3 \varepsilon G_8) = 0.2$, and $P(H_2 \varepsilon G_{10}) = 0$. Therefore, for a given group G_q, a measure of expected distinctiveness can be defined as

$$E(G_q) = \sum_{i=1}^{n} s_{iq}\ P(H_i \varepsilon G_q) = \sum_{i=1}^{n} s_{iq}\ P(s_{iq})$$
$$= \sum_{i=1}^{n} s_{iq}\ \frac{d(G_q)\ s_{iq}}{D}.$$

For a given partition P_ℓ, defined by a vertical line through class level ℓ, the expected level of distinctiveness $E(P_\ell)$ for this partition is then defined as

$$E(P_\ell) = \sum_{q \varepsilon P_\ell} \sum_{i=1}^{n} s_{iq} \left(\frac{d(G_q)\ s_{iq}}{D} \right)$$

$$= \frac{1}{D} \sum_{q \epsilon P_\ell} \left[d(G_p) \sum_{i=1}^{n} (s_{iq})^2 \right]. \qquad (3.11)$$

Since s_{iq} is a (0;1) variable, $s_{iq} = (s_{iq})^2$, and $\sum_{i=1}^{n} (s_{iq})^2$ simply equals the number of hospitals in group G_q. Thus, the measure of expected distinctiveness for any given partition P_ℓ, (3.11), is simply a constant term $\left(\frac{1}{D}\right)$ times the sum of the number of hospitals in each group in partition P_ℓ times the distinctiveness (in terms of class levels) of that group. Since the term $\left(\frac{1}{D}\right)$ has no effect (D is constant for our computer programs), the term may be dropped and the expected distinctiveness E(P) for partition P hereafter defined as

$$E(P) = \sum_{q \epsilon P} \left[d(G_q) \sum_{i=1}^{n} (s_{iq}) \right] . \qquad (3.12)$$

The expected distinctiveness values for each group are shown in Figure 3.8 (overlay #2); summing these values for all groups defined for each partition gives an objective measure for evaluating the partition formed at each class level. As apparent from Figure 3.8, the partition formed at the fourth class level in this example appears to be optimal with $E(P_\ell) = 30$.

B. Optimal Partition Determination

Given that any partition P_ℓ can be evaluated by its expected distinctiveness E(P), the problem now is to find the partition P which maximizes $E(P_k)$. The problem is complicated by the fact that the optimal partition isn't necessarily defined by a vertical line through any class levels; thus, evaluating $E(P_\ell)$ for all values of $\ell = 1, 2, \ldots, 25$ will not guarantee identification of the optimal partition.

In order to find the optimal partition, as well as suboptimal partitions and the tradeoffs between partitions and expected distinctiveness, define decision variables y_q as follows:

$$y_q = \begin{cases} 1 & \text{if group } G_q \text{ is included in the optimal partition } P^*, \\ 0 & \text{otherwise} \end{cases}$$

Then, letting $c_q = d(G_q) \sum_{i=1}^{n} s_{iq}$, the problem of finding the optimal partition P* can be formulated as an integer linear programming problem below:

(SPP) Maximize $E(P_k) = \frac{1}{D} \sum_{q=1}^{n} c_q y_q$

S.T.

$$\sum_{q=1}^{n} s_{iq} y_q = 1 \quad \text{for all } i = 1, 2, \ldots, n$$

$$y_q = 0, 1 \quad \text{for all } q = 1, 2, \ldots, M.$$

where

M = total number of identified groups.

Using the example shown in Figure 3.8 to illustrate the problem where the c_q values are indicated in the second overlay and M = 11 groups, the problem of finding P* is stated as follows:

$$\text{Maximize } E(P) = 4y_1 + y_2 + y_3 + 2y_4 + 5y_5 + 5y_6 + 2y_7 + 6y_8 + 20y_9 + 9y_{10} + 6y_{11}$$

S.T.

$$\begin{aligned}
y_1 &+ y_9 + y_{11} = 1 \\
y_2 &+ y_7 + y_8 + y_9 + y_{11} = 1 \\
y_3 &+ y_7 + y_8 + y_9 + y_{11} = 1 \\
y_4 &+ y_8 + y_9 + y_{11} = 1 \\
y_5 &+ y_{10} + y_{11} = 1 \\
y_6 &+ y_{10} + y_{11} = 1 \\
& y_q = 0,1
\end{aligned}$$

Note that the term $\left(\frac{1}{D}\right)$ was dropped as this constant term would not affect the optimal solution determination in any way.

Problem (SSP) is a well known and widely studied combinatorial problem referred to as the set partitioning problem (see Garfinkle and Nemhauser, 1972). Theoretically, problem (SSP) could be solved by a general purpose integer linear programming (IP) algorithm; however, even state-of-the-art IP algorithms would severely limit the number of variables and hence the problem size which could be accommodated. While a number of specialized algorithms have been proposed for the set partitioning problem, it remains, in general, a difficult problem to solve. However, the presence of a single property in this case, based on the nature of a dendrogram, suggests a solution approach which is easily implemented and guarantees optimality.

The solution procedure is based upon the observation that a dendrogram is simply a specific type of graph called a tree, where the vertical line segments and terminal points in a dendrogram can be considered nodes (G_q) and horizontal line segments can be considered directed arcs with weights c_q (Figure 3.8). A tree then is a finite graph without a cycle and with at least two vertices. If every node except one (node G_{11}) of a tree is the terminal node of exactly one arc, then the tree is said to be an arborescence of root G_{11}.

The algorithm begins at the root of the arborescence G and searches through lower cluster levels; the solution procedure is based upon the observation that the branches of the tree automatically partition set H. Letting I_j be the set of j arc indices branching from node G_j, problem (SPP) becomes one of finding sets I_j such that $\sum_{j=1}^{J} \sum_{q \in I_j} c_q$ is maximized (where c_q -- the expected distinctiveness of group G_q -- represents the weight on the arc terminating at node G_q). The nature of the dendrogram guarantees that the constraints are met (i.e., that $\sum_{j=1}^{J} \sum_{q \in I_j} G_q$ partitions H); the objective function allows the problem to be decomposed into J subproblems.

The algorithm begins at node G_M and searches the branching from nodes G_M; each of these nodes is subsequently examined until the algorithm reaches the terminal nodes of the tree (i.e., hospitals grouped individually) or until certain conditions are met. Assuming that the search procedure is at node G_j, where G_j is not a terminal node of the tree, the following three rules establish conditions for continuing or stopping the search procedure along any branch, and guarantee that a global optimum will be found.

Rule 1: Find the set of arcs I_j branching from node G_j and calculate $\sum_{q \in I_j} c_q$. If $\sum_{q \in I_j} c_q > c_j$ (where c_j represents the weight on the arc terminating at node G_j), then the search from all nodes $G_r (r \in I_j)$ must be continued and group G_j eliminated from consideration.

Justification: If $\sum_{q \in I_j} c_q > c_j$, the weights of the branches emanating from G_j are greater than the weight on branch c_j. Since problem (SPP) can be decomposed into independent subproblems, only the set of nodes I_j need be considered.

Rule 2: Calculate $\max_j = \sum_{i=1}^{n} s_{ij} \left[\ell(G_j) \right]$ at node G_j, where $\ell(G_j)$ is the class level corresponding to node G_j. If $\max_j > c_j$, then search is halted from node G_j and $G_j \in P^*$.

Justification: Since $\sum_{i=1}^{n} s_{ij}$ = number of hospitals in set I_j, the maximum possible expected distinctiveness value from node G_j would be defined by (3.11) if $d(G_j) = \ell(G_j)$. Thus, if $c_j > \max_j$, no improvement is possible, any search from node G_j is

unnecessary, and $G_j \varepsilon P^*$.

Rule 3: At node G_j, find $\min_j$ (the minimum distinctiveness value possible from node G_j). If $\min_j > c_j$, then the search must continue and $G_j \varepsilon P^*$.

Justification: Since $\min_j$ represents the worst possible case from node G_j, if $\min_j > c_j$, then continuing the search must improve the total distinctiveness value.

The calculation of $\min_j$ is based upon the class level corresponding to node G_j $\left(\ell(G_j)\right)$, the maximum number of subsequent partitions (K_j) possible from node G_j, and the number of hospitals (n_j) in group G_j (where $n_j = \sum_{i=1}^{n} s_{ij}$). The value of K_j is based upon finding the maximum number of possible groups from node G_j which occurs if all subsequent groups are formed by pairwise combinations. Given that there are, on the average, $\frac{n}{2^k}$ groups at the k^{th} class joining, the maximum number of groups is defined as follows,

$$\sum_{k=0}^{K_j} \frac{n_j}{s^k} + \left(\frac{n_j}{2^{K_j}} - 1 \right) = n_j \left(2 - \frac{1}{2}^{K_j - 1} \right) + \left(\frac{n_j}{2^{K_j}} - 1 \right) = 2n_j - 1,$$

where K_j is the largest value such that $\frac{n_j}{2^{K_j}} > 1$. (This problem is equivalent to the problem of finding the number of nodes in a binary tree.) Given the $(2n_j - 1)$ groups, the problem of finding $\min_j$ is stated as follows:

$$\text{Minimize} \left\{ \text{Maximum}_{i=1,\ldots,K_j} \left(\sum_{n \varepsilon P_{ij}} E(G_h) \right) \right\} \tag{3.13}$$

$$\sum_{h=1}^{2j_j - 1} E(G_h) = n_j \, \ell(G_j)$$

$$E(G_L) \geq 0 \quad \text{for all } L$$

where a dendrogram of all pairwise combinations from node G_j defines partitions P_{ij}. Since the number of hospitals in each group G_h can be easily determined, the problem of defining $E(G_h)$ requires determining $d(G_h)$. Problem (3.13) can be simplified somewhat, however, by recognizing that all group joinings in any partition must take place at the same level in order to maximize $\sum_{h \varepsilon P_{ij}} E(G_h)$ for any partition P_{ij}. Thus, problem (3.13) can be reduced to finding distances for any partition P_{ij}.

Unfortunately, not all groups will join in a given partition if there exists an odd number of groups; thus, the distance (and expected distinctiveness) for this group will extend into the next partition. For example, consider the dendrogram of pairwise joinings in Figure 3.11 where $K_j = 3$, $\ell(G_j) = 7$, and $n_j = 5$. In this case, group G_5 in partition P_{1j} cannot join another group until level 5, thereby increasing $E(P_{2j})$ from 15 to 17 (the maximum distinctiveness in this case).

Therefore, given the distances $d(P_{ij})$ for partitions P_{ij} (where partitions are defined by group joinings), the maximum possible expected distinctiveness value is defined as

$$\max_{i=1,\ldots,K_j} \left\{ n_j \; d(P_{ij}) + c_i \left[d(\cdot) \right] \right.$$

where

$$c_i \left[d(\cdot) \right] = (2^{i-1}) \; d(P_{i+1,j}) \; 6_{ij} + (2^{i-2}) \; d(P_{i-1,j}) \; y_{i-1,j}$$

and

$$y_{ij} = \begin{cases} 1 & \text{if } \left\langle \frac{n}{2^{i-1}} \right\rangle \text{ is odd,} \\ 0 & \text{otherwise,} \end{cases}$$

and $\langle \cdot \rangle$ represents the smallest integer greater than or equal to $(\cdot)$. (Calculations are represented in Table 3.4; for example, if $\langle \frac{n}{2} \rangle$ and $\langle \frac{n}{3} \rangle$ are both odd, then $E(P_{3j}) = n_j d(P_{3j}) + 4 \; d(P_{4j}) + 2 \; d(P_{2j})$.)

TABLE 3.4

Minimum Expected Distinctiveness Calculation

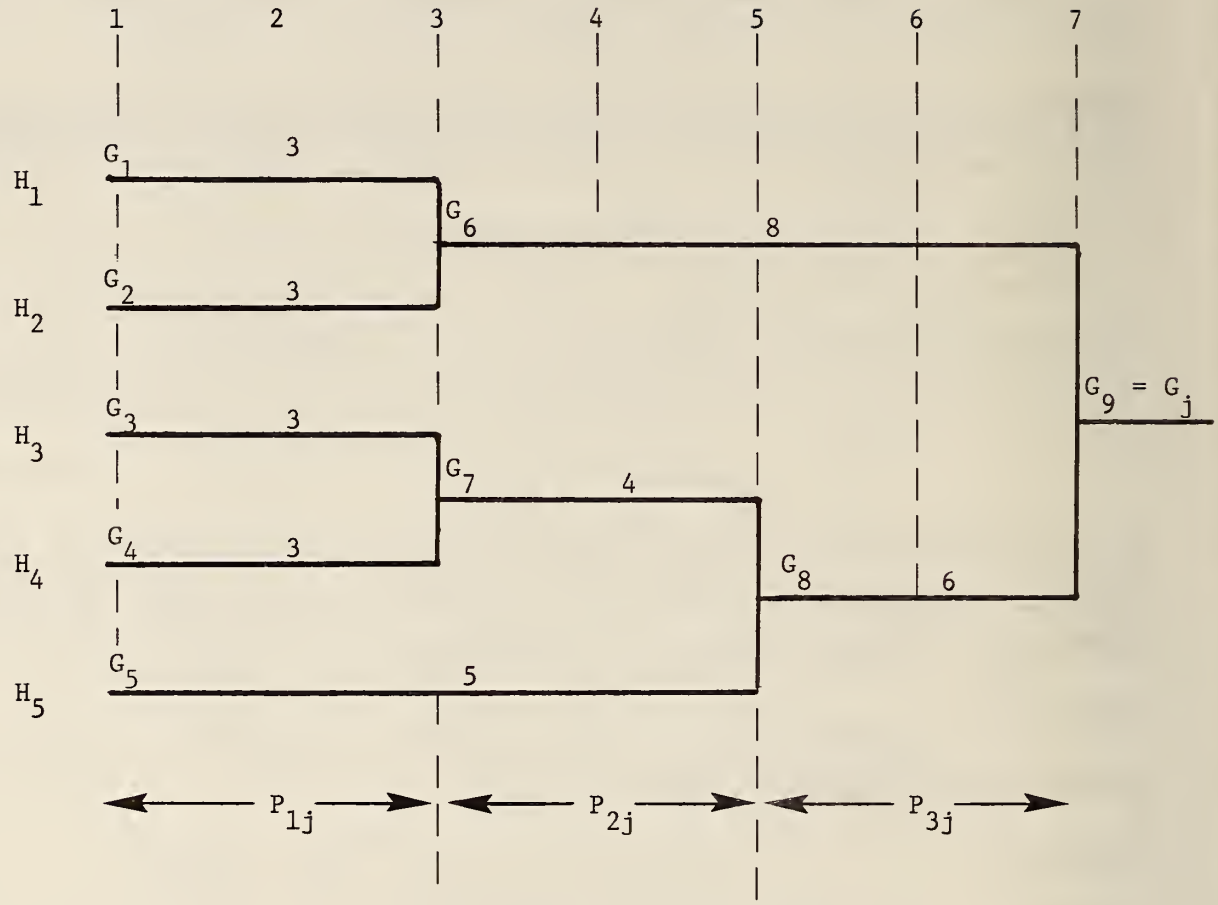
PARTITION	1	2	3	...	i	...	K
Distance:	$d(P_{1j})$	$d(P_{2j})$	$d(P_{3j})$	...	$d(P_{ij})$	...	$d(P_{K_j})$
No. of groups:	n	$\langle \frac{n}{2} \rangle$	$\langle \frac{n}{3} \rangle$	...	$\langle \frac{n}{2^{i-1}} \rangle$	...	$\langle \frac{n}{2^{K_j-1}} \rangle$
No. of hospitals in merged groups:	1	2	3	...	2^{i-1}	...	2^{K_j-1}

The problem of finding $\min_j$ then can be stated as

$$\text{Min} \left\{ \max_{i=1,\ldots K_j} (n_j d(P_{ij}) + c_i \left[d(\cdot) \right]) \right\} \qquad (3.14)$$

FIGURE 3.11

Binary Dendrogram Illustration



S.T.

$$\sum_{i=1}^{K_j} d(P_{ij}) = \ell(G_j)$$

$$d(P_{ij}) \geq 0$$

While problem (3.14) is difficult to solve in general, it was noted that in most cases the optimal solution to problem (3.14) could be found by setting $d(P_{1j}) = \left\langle \frac{\ell(G_j)}{K_j} \right\rangle$, $d(P_{2j}) = \left\langle \frac{\ell(G_j) - d(P_{2j})}{K_j - 1} \right\rangle$, $d(P_{3j}) = \left\langle \frac{\ell(G_j) - d(P_{2j}) - d(P_{3j})}{K_j - 2} \right\rangle$, etc. In addition, the computer programs used to construct the dendrograms (Anderberg, 1973) arbitrarily use 25 class levels--the upper bound for $\ell(G_j)$. Thus, tables displaying values of $\min_j$ could be easily constructed for values of $\ell(G_j)$ and various group sizes (n_j).

The entire algorithm for finding P* was never computerized as it was found that the search procedure effectively allowed manual solutions to be found in a few minutes, even for the largest data set (1,070 hospitals) and most complex (i.e., sensitive) composite dendrograms.

Not only can partitions be easily found and evaluated by this approach, but a tradeoff between the number of groups and total expected distinctiveness can be evaluated. To examine suboptimal partitions, it is simply necessary to find which groups should be split or combined to make the smallest marginal decrease in E(P*). This is easily accomplished by arbitrarily setting $c_q = \infty$ for each $q \varepsilon P^*$ and resolving problem (SPP). A graph showing the tradeoffs between the number of groups and expected distinctiveness for the example in Figure 3.8 is shown in Figure 3.12.

V. Criteria for Cluster Validation

Testing the validity of the hospital partitions is essential if credibility is to be placed on any reimbursement system derived from the resultant groups. While there is not single test or procedure for measuring "goodness of fit" in multivariate grouping problems, a number of approaches were adopted in this study which allowed for a through examination of clustering results. The approaches used here, which are briefly described below, fall into three major categories: (1) descriptive statistics for examining the reasonableness, intuitive consistency, and parsimony of the identified groups, (2) nonparametric tests for statistical consistency, and (3) parametric tests based on multivariate normal assumptions.

A. Descriptive Statistics

Examination of the identified groups initially consisted of visual group inspection, including nonquantifiable group characteristics such as state and county

names, types and hospitals, etc., and the calculation of the mean, standard deviation, and measures of skewness and kurtosis for each independent characteristic of the group. On the basis of the similarity measure used (the squared Euclidean Distance), several statistics were found, including the within group diameters (i.e., the maximum within group distance), the within group sum of squared distances, and the distance to each group centroid.

B. Nonparametric Tests for Statistical Consistency

Most proposed tests of cluster validation may themselves be classified into two categories: (1) tests based upon results from various clustering schemes, and (2) tests based upon statistics calculated from multivariate normal assumptions. It should be noted that tests comparing results from various clustering schemes are relative tests; i.e., they can only indicate differences among alternative clustering techniques but not the direction of those differences. Tests assuming that the data have been drawn from distinct populations of known structure, on the other hand, can measure the absolute effectiveness of resultant paratitions. Both types of tests are profitable and were used in this study.

A problem frequently encountered throughout this study concerned measuring the degree of similarity (or dissimilarity) between two or more partitions. For example, it was necessary to compare groups of hospitals determined by our cluster analytic approach with groups used by HCFA which had been determined by multiple cross classification (described in the following chapter).

Recently, Rand (1971) suggested a statistic for measuring the similarity between partitions of hospitals based on three assumptions: (1) each hospital is uniquely assigned to a subset or group (i.e., no overlapping groups are considered), (2) it is equally meaningful for a partition to not group two hospitals together as it is for a partition to group the hospitals, and (3) all hospitals are equally weighted in group determination. Furthermore, Rand demonstrated several properties and simplified computational forms for the statistic.

Unfortunately, it is possible to show that Rand's suggested statistic which is equivalent to a simple matching coefficient used to describe similarity between vectors of binary values, may be seriously distorted by the second assumption above when relatively large numbers of nontrivial groups exist in a partition. Therefore, a new statistic was calculated which eliminated any distortion.

Given two hospitals, each described by a vector of m dichotomous characteristics $[x_{ij}, \ldots, x_{cj}, \ldots, x_{mj}]^T$, where

$$x_{cj} = \begin{cases} 1 \text{ if the } j^{th} \text{ hospital contains the } c^{th} \text{ characteristic} \\ 0 \text{ otherwise} \end{cases}$$

a number of measures (usually known as coefficients of association) can be constructed which measure the overall similarity between the i^{th} and the j^{th} hospital. One measure, s(j,j), introduced by Sokal and Michener (1958) and studied by Goodall (1967), is known as the simple matching coefficient and is simply defined as the ratio of the number of matches to the total number of

characteristics; or, more precisely, if

$$\lambda_c = \begin{cases} 1 \text{ if } x_{ci} = x_{cj} \\ 0 \text{ otherwise} \end{cases} \tag{3.15}$$

then

$$s(i,j) = \sum_{c=1}^{m} \frac{\lambda_c}{m} \tag{3.16}$$

This simple matching coefficient (3.16) is often challenged on the grounds that mutually characteristic absences (i.e., zeroes in the characteristic vectors for both hospitals) contribute to the measure of similarity. Therefore, an opposite approach is to simply eliminate the counting of these mutually lacking characteristics. This measure, $\bar{s}(i,j)$, known as the Jaccard coefficient, is more precisely expressed by redefining λ_c, call it $\bar{\lambda}_c$, as,

$$\bar{\lambda}_c = \begin{cases} 1 \text{ if } x_{ci} = 1 \text{ and } x_{cj} = 1 \\ 0 \text{ otherwise} \end{cases}$$

and defining $\bar{s}(i,j)$ as follows:

$$\bar{s}(i,j) = \sum_{c=1}^{m} \frac{\lambda_c}{m - \left(\sum_{c=1}^{m} \lambda_c - \sum_{c=1}^{m} \bar{\lambda}_c \right)} . \tag{3.17}$$

(A myriad of additional measures exist on the continuum between these two measures, which mostly use various schemes for weighting matches and mismatches. For a description of fourteen such measures, see Anderberg, 1973.)

Given n hospitals segregated by two partitions P_i and P_j, measures describing a degree of similarity between the two partitions can be constructed by treating pairs of hospitals as characteristics and proceeding in a manner analogous to the measures of the preceeding section. Letting $C = \{1, 2, \ldots, \binom{n}{2}\}$ be the set of all hospital pairs, a measure equivalent of the simple matching coefficient can be derived by letting

$$x_{ck} = \begin{cases} 1 & \text{if partition } P_k \text{ groups the } c^{th} \text{ hospital pair together} \\ 0 & \text{otherwise} \end{cases}$$

defining λ_c by (3.15), and summing (3.16) over all hospital pairs; i.e.,

$$s(i,j) = \sum_{c \in C} \frac{\lambda_c}{\binom{n}{2}} , \tag{3.18}$$

The simple coefficient (3.18) was proposed by Rand (1971); the Jaccard coefficient equivalent for partitions can be defined in similar fashion using λ_c and (3.17).

The simple matching coefficient has the disadvantage that when comparing two partitions with relatively large numbers of (nontrivial) groups, its value will tend to increase directly as a function of the number of groups, due to the fact that hospital pairs tend to spread across groups and thus increase the number of "nonmatches." Alternatively, consider the hospital being randomly partitioned. As the number of groups increase, the simple matching coefficient fails to account for the increasing probability of two hospitals not being grouped together. Conversely, the Jaccard equivalent would present exactly opposing difficulties, as it would fail to consider the higher likelihood that two hospitals will be grouped together when fewer groups exist. Thus, it would appear that a measure, adjusted for the number of groups in each partition, should have values between those of the simple matching coefficient and the Jaccard coefficient.

To develop a measure less dependent on the number of groups present, assume that partition P_i has n_i groups and hospitals are grouped randomly.[42] Then, the probability that the c^{th} pair of hospitals will be grouped together in partition P_i is simply $P_r(x_{ci} = 1) = \frac{1}{n_i} \frac{1}{n_j}$. The probabilities for the four possible cases for the c^{th} hospital pair are summarized in Table 3.5.

TABLE 3.5

Probabilities for Random Grouping of the c^{th} Hospital Pair

		Partition P_i	
		$P_r(x_{ci} = 0)$	$P_r(x_{ci} = 1)$
Partition P_j	$P_r(x_{cj} = 0)$	$(1 - 1/n_i)(1 - 1/n_j)$	$(1/n_i)(1 - 1/n_j)$
	$P_r(x_{cj} = 1)$	$(1 - 1/n_i)(1/n_j)$	$(1/n_i)(1/n_j)$

Then, using the probabilities that these events do not occur randomly, an undistorted statistic Λ is easily defined as follows:

$$\Lambda = \frac{\sum_{c \in C} \lambda_c \left|1 - P_r(x_{ci})\, P_r(x_{cj})\right|}{\sum_{c \in C} \left|1 - P_r(x_{ci})\, P_r(x_{cj})\right|} \qquad (3.19)$$

where the probabilities shown in Table 3.4 are used as weights and the matching parameter λ_c is defined by (3.15).

---

[42] With this assumption, the development of this statistic is conceptually similar to a measure of association by Hyvärinen (1962).

To illustrate the calculation of Λ, assume that five hospitals are partitioned into two partitions, P_1 = {A} {BCD} {E}, and P_2 = {AB} {CDE}. Examining all possible hospital pairs, Table 3.6 indicates the ten values of x_{ci}, λ_c, and $P_r(x_{ci})$ for each pair.

TABLE 3.6

Numerical Illustration - Five Hospitals

Weight Calculation	Hospital Pair AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
					Partition P_1					
x_{c1}	0	0	0	0	1	1	0	1	0	0
$P_r(x_{c1})$	.67	.67	.67	.67	.33	.33	.67	.33	.67	.67
					Partition P_2					
x_{c2}	1	0	0	0	0	0	0	1	1	1
$P_r(x_{c2})$	.5	.5	.5	.5	.5	.5	.5	.5	.5	.5
					Weights					
λ_c	0	1	1	1	0	0	1	1	0	0
$1-P_r(x_{c1})P_r(x_{c2})$	.667	.667	.667	.667	.834	.834	.667	.834	.667	.667

In this example, the simple matching coefficient (3.18) gives a value of 0.5 and the Jaccard coefficient gives a value of 0.167; the Λ coefficient equals 0.488.

The measure Λ defined above varies from 0 (no pairwise matching) to 1 (complete pairwise agreement between partitions). If all probabilities $P(x_{ci})$ are equal (the case when only two groups are present), Λ reduces to the simple matching coefficient.

In order to test the Λ statistic twenty hospitals were randomly partitioned into varying numbers of groups by randomly assigning each hospital a value

from 2 to K (where K is the number of groups). As the value of K increased, the proportion of non-empty groups decreased substantially, as evident in Table 3.7. After randomly determining two partitions, the three statistics indicated were calculated for forty trials and averaged. The results are displayed in Table 3.7--as expected, only the proposed statistic consistently indicates that approximately one half of the hospitals are similarly grouped.

If groups are randomly determined, the expected value of Λ if 0.5. By realizing that the expected number of OTU pairs in each event listed in Table 3.5 is simply $P(x_{ci})\ P(x_{cj})\ \binom{n}{2}\ 1-P(x_{cj})$, it is easily shown that the expected value of Λ reduces to

$$E(\Lambda) = \frac{2\ \binom{n}{2}}{4\ \binom{n}{2}} = 0.5.$$

The undistorted statistic developed here was widely used in this study. For example, the statistic was used to compare optimal and suboptimal partitions, partitions determined in this study and the partitions created in SSA cross classification scheme, partitions created by discriminant functions, etc.

C. Parametric Tests for Statistical Validity

The preceeding sections have alluded to the use of parametric tests for examining the validity of resultant clusters. If it is assumed that a given number of populations exists and it is the task of the clustering procedures to correctly identify those populations strong tests of significance can be developed for measuring absolute effectiveness of a clustering result.

Parametric tests are based on the assumption (supported for large samples by the Central Limit Theorem) that elements in each population are distributed by a multivariate normal density function. Accepting this assumption, three approaches were used to test the resultant clusters: (1) one-way analysis of variance (ANOVA), (2) linear discriminant analysis, and (3) regression analysis.

C.1. Analysis of Variance

In order to test resultant clusters, a one-way analysis of variance was used to test for significant differences between all identified groups for all variables (i.e., factor scores). That is, if seven factors were used to describe each hospital, seven one-way analyses of variance were run using each factor as the independent variable and a variable indicating group as the dependent variable. Based on the sum of squared deviations between groups and the sum of squared deviations not explained by the independent variable, an F ratio was calculated and used to test if significant differences exist between groups for each factor used. For these tests (as well as other parametric tests), a significance level of 0.05 was used.

TABLE 3.7

Comparisons of 20 Ramdomly Partitioned Hospitals*

Number of Groups	Average Number of Nonempty Groups	Simple Matching Coefficient	Jaccard Coefficient	Λ
2	2.00	.495	.337	.495
3	3.00	.552	.194	.496
4	4.00	.619	.139	.494
5	4.95	.679	.121	.502
6	5.84	.716	.088	.493
7	6.67	.759	.073	.505
8	7.45	.782	.069	.504
9	8.17	.804	.051	.502
10	8.95	.826	.051	.513
11	9.42	.834	.051	.504
12	9.95	.849	.053	.510
13	10.46	.863	.043	.513
14	10.84	.868	.029	.503
15	11.21	.879	.040	.513
16	11.67	.886	.029	.509
17	12.12	.893	.027	.514
18	12.27	.897	.027	.510
19	12.69	.902	.030	.509
20	12.55	.901	.025	.494

*All results are based on simulation runs of 40 trials for each target group size.

C.2. Discriminant Analysis

Discriminant analysis, previously mentioned in the first chapter, assumes that groups have been identified and attempts to construct linear functions which explain the differences between groups. Thus, having found clusters, we assumed that each cluster identified distinct populations normally distributed about their respective cluster centroids, and employed linear discriminant analysis to find these functions and reclassify the hospitals. In this case, information on the accuracy of the discriminant functions is provided by the number of hospitals which are correctly reclassified by the functions, and Wilks lambda, which indicates whether or not the functions are statistically significant.

If significant, discriminant analysis offers one means for reclassifying uniquely grouped hospitals (i.e., isolates). Using the discriminant functions to reclassify isolates in this case is equivalent to computing the squared Euclidean distance to the centroid of the groups determined by the discriminant analysis, computing each isolate's respective chi squared value from this distance, finding the probability of this chi squared value (from a chi square table)--where this quantity determines the proportion of hospitals which lie as close to or closer to the cluster centroid, and assigning the isolate to the group which maximizes this probability of membership.

C.3. Regression Analysis

If we assume that variations in cost per case are a function of variation in both the economically "legitimate" variables (identified in Chapter Two) which were used to determine the clusters and "non-legitimate" or inefficiency variables (e.g., number of beds, number of RN's and LPN's, etc.), and if clusters reduce the amount of variation in legitimate variables (as tested by the ANOVA), then it has been argued that it should be possible to detect an appropriate reduction in cost per case.[43] One means for testing this hypothesis is provided by regression analysis; i.e., using cost per case as the dependent variable and both legitimate and inefficiency variables as the independent variables, find the appropriate regression equations for each identified cluster. If the hypothesis stated above is true, then the regression coefficients for the inefficiency variables should explain the majority of variation in cost per case and the legitimate variables should have little or no significant effect (for each cluster).

One would expect to find this difference between regression coefficients, however, only if the following two assumptions are valid: (1) the functions relating cost per case to the independent variables must be linear, and (2) the points representing hospitals must in fact approach a multivariate normal swarm around each cluster centroid. The uncertainty associated with the first assumption was a major factor contributing to the selection of cluster analysis, as discussed in the first chapter. The second assumption is basic for any parametric test, but again there is no a priori reason to expect this to hold, especially in clusters with relatively fewer hospitals. The importance of this assumption can be illustrated by the example represented in Figure 3.4; assume that in a final partition, hospitals H_2 and H_3 in this example, are grouped

---

[43] For further discussion of this point, see Phillip and Iyer (1972).

together (a reasonable assumption as the distance between the points representing these two hospitals is relatively small). However, if one determines a regression equation using income (I) as well as inefficiency variables as independent variables, it is entirely possible that the income variable could explain 100 percent of the variation in cost per case between the two hospitals. While this example of two hospitals is, admittedly, extreme, it underscores the importance of assuming clusters represent multivariate normal swarms around each cluster centroid.

In addition, it is assumed that variation in the cost per case (dependent variable) not be dominated by the inefficiency variable. If this were to occur, one would still detect significant differences between regression coefficients and coefficients of determination within clusters, but would not be able to detect any significant differences between clusters.

Therefore, while the results of such regression analyses are reported in the following chapter, it is exceedingly important that these results be most cautiously interpreted and carefully used if no significant differences are found.

CHAPTER FOUR

EMPIRICAL RESULTS

I. Introduction

In order to illustrate the economic framework and test the feasibility of the cluster analysis methodology, the concepts developed in Chapters Two and Three were applied to a 1973 data base compiled by the Office of Policy, Planning, and Research, Health Care Financing Administration (HCFA). The data set contains information gathered from Medicare cost reports for 1973, Census Bureau reports (1970 census), the American Hospital Association (AHA) annual survey of hospitals for 1973, and a variety of demographic and socio-economic data files for 1097 short-term, general U. S. hospitals.

From the outset, it was clear that the nature of the available data set precluded anything other than a preliminary test of the feasibility of our classification model. Direct measures for a number of the variables identified in Chapter Two were not available (e.g., case mix and case mix severity), and there were no acceptable means of evaluating the bias introduced by substituting the surrogate measures that were available. Thus, the empirical results reported here must be viewed as suggestive rather than conclusive. They are of value only in that they provide useful and important information about the feasibility of the approach, potential problem areas, and directions for further investigation.

Given the problems of data availability, the first step in the empirical analysis was to select measures that would be used to reflect the impact of the variables identified in Chapter Two: input prices, rural markets, case mix, and case mix severity. The measures used for the input prices and rural markets variables are described in Chapter Two. Ideally, diagnostic data would be used for the latter two variables.

However, as indicated above, no direct diagnostic data were available. In their place, two alternative approaches were used. In the first, the effects of variations in case mix were represented by a number of measures based on endogenous characteristics of the individual hospitals. This approach was labeled the endogenous approach. In the second, a number of measures reflecting the socioeconomic character of the county in which the hospital is located were used. Since these measures are exogenous to the individual hospital, this method was called the exogenous approach. The impact of differences in case mix severity was captured by adding a number of demographic variables in the endogenous approach. No further variables were needed for case mix severity in the exogenous approach. A complete listing of the two sets of measures is given in Table 2.3.

For comparison, a third approach was also used. The measures in this approach were based on the cross classification system previously used by HCFA to group hospitals that provide services to Medicare beneficiaries. The HCFA variables consisted of the following: (1) median family income in the state, (2) bed

size of the hospital (at the end of 1973), and (3) a dummy variable indicating whether or not the hospital is located in an SMSA.[44]

In order to remove implicit weights caused by multicollinearities among the measures, all sets of measures were first factor analyzed and subsequently represented only by factor scores (as discussed in the third chapter). Two types of explicit weighting schemes were then applied to the three sets of factor scores. The first type was based on the standardized linear regression coefficients associated with the factors, estimated using cost per case as the dependent variable. The second set of weights was determined in such a way that each variable was given equal (unit) weight. These steps are discussed in more detail below.

In order to refine the methodology and test the feasibility of using a small sample to estimate groups for the entire population, a random subsample of 194 hospitals was selected. The smaller sample size made it possible to easily analyze all sets of measures and weights, consider a number of analytic possibilities, and fully specify composite dendrograms. Following analysis of the subsample, the complete sample was analyzed and the results were compared.

A. Data Base Description

The data base was initially examined for any missing and/or questionable values; missing values were identified and replaced from secondary data sources whenever possible. Questionable values were defined as those values lying outside a three standard deviation interval about the variable mean. Any hospital having questionable values which could not be verified was deleted from the sample.

Of the original 1097 hospital sample provided by HCFA, a verified sample of 1070 hospitals remained after screening. Of the hospitals remaining, 86 percent were accredited by the JCAH, 29 percent of the hospitals had some kind of medical school affiliation, and 60 percent were located within an SMSA. The average bed size was 282.4 beds, with a standard deviation of 241 beds and a skewness of +1.26 (indicating that the mode or most frequently occuring value is less than the average value). The measures used to represent the relevant classification variables and their mean values among sample hospitals one given below.

As indicated in the second chapter, input factor prices are represented by three county based wage measures and one hospital based wage measure. These measures

---

[44] The HCFA system places a hospital into a given group based on its number of beds (0-54 beds, 55-94 beds, 100-169 beds, 170-264 beds, 265-404 beds, 405-684 beds, and over 685 beds), whether or not the hospital is in an SMSA, and a state-based per capita income category. The number of variable categories used (7 for bed size, 2 for SMSA/non-SMSA, and 5 for per capita income) results in 70 possible cross classification cells for the HCFA system, which is supposed to represent all reasonable combinations of hospital size and economic environment.

and their sample means are as follows:

Manufacturing wages (hourly)	$4.16
Transportation & public utilities wages (hourly)	$4.63
Retail wages (hourly)	$2.34
Hospital wages (annual)	$7365.28

The rural markets variable is represented by the uniform pressure occupancy index (UPOI) described in section IV, part D, Chapter Two. The mean value of this index for the complete hospital sample is 89 percent.

The endogenous approach uses a number of hospital specific measures to represent case mix and case mix severity. These case mix surrogate measures and their mean values are listed below.

Number of basic services (maximum value of 4)	3.9
Number of quality enhancing services (maximum value of 7)	4.6
Number of complex services (maximum value of 16)	7.5
Number of community services (maximum value of 17)	4.4
Number of births per discharge	0.095
Number of surgical operations per discharge	0.19
Number of outpatient visits per discharge	4.97

To measure case mix severity, the percentage of the county's population under 5 years old and the percentage of the population over 65 years were combined under the assumption that those two age groups produce the most severe cases. A second severity measure, the percentage of families earning less than $4,000 yearly, assumes that poverty families are more likely to be treated for more severe cases. The mean values for these measures are as follows:

Percentage of population under 5 and over 65	21%
Percentage of families earning less than $4000 yearly	17%

The exogenous approach used a number of county specific measures to represent case mix and case mix severity. One of these measures, "heavy demand ages," was determined by adding the percentage of the population under 5, the percentage of population over 65, and the percentage of females from 15 to 44 years old. The mean values for all exogenous measures are listed below.

Heavy demand age	41.3%
Percentage of families earning less than $4000 yearly	17.0%
Family income (yearly)	$9146.29
Labor force participation rate	56.3%
Disabled rate ages 16-44	10.1%
Percentage of population non-white	10.9%
Percentage of M.D.'s aged 60 and over	20.9%
Obstetricians/Gynecologists per 10,000 population	0.0833
Primary care M.D.'s per 10,000 population	0.8134
Medical specialists per 10,000 population	3.1877
Direct care specialists per 10,000 population	0.2699
Other specialists per 10,000 population	0.2196
Surgical specialists per 10,000 population	0.4062

The three variables used in the third approach based on the HCFA cross classification system were median family income by county, total number of hospital beds, and a dichotomous zero-one variable indicating whether or not the hospital was located within an SMSA. While the HCFA system uses statewide average family income per state, such data were not available in this study. Hence, the county-based measure was substituted. The three measures and their mean values are given below.

Median family income (by county)	$9146.30
Total number of hospital beds	282.40
Percentage of hospitals within a SMSA	60.40%

As part of the initial examination of the data base, the simple product moment correlation coefficient between all pairs of measures was computed. The correlation coefficient is of interest since it indicates the degree of multicollinearity and hence the implicit weighting present among the measures. We found, for example, that many of the proposed case mix measures were highly correlated with input price measures. Failure to remove this multicollinearity would result in implicit increases in the weights on both the input price and case mix variables. Thus, it became imperative to factor analyze the measures in order to attempt to isolate the parts of the measures corresponding to each of the dimensions.[45]

A subset of these correlation coefficients is given in Table 4.1.; the complete correlation matrices are given in Appendix B. Due to the large sample size, coefficients with absolute values as low as 0.20 are statistically significant.

B. Factor Analysis

Given the relatively high correlations found among measures, it was necessary to factor analyze each set of measures in order to eliminate the implicit variable weighting caused by these multicollinearities. In essence, the factor analysis simply reduced each data set to strictly orthogonal measures.

Each factor analysis was computed using varimax rotation to simplify the rows (i.e., variable loadings) as much as possible, and an eigenvalue limit of 1.0 (the default value) was used.[46] In all cases, this limit resulted in factors which accounted for a minimum of 80 percent of the total data set variation.

The first set of measures to be factor analyzed (the endogenous approach measures) resulted in six factors which accounted for 80.3 percent of the total variation.

---

[45] Note that in the unlikely case that all measures are completely orthogonal, there would exist as many factors as measures and the resultant similarity scores would be the same if computed from factor scores or from the standardized measures themselves.

[46] The SPSS (Statistical Package for the Social Sciences) program was used for the regression and factor analyses.

TABLE 4.1

Product - Moment Correlation Coefficients

Endogenous Measures

Manufacturing wages with percentage of families earning less than $4,000 yearly	-0.69
Transportation & public utilities wages with percentage of families earning less than $4,000	-0.61
Retail wages with percentage of families earning less than $4,000 yearly	-0.56
Transportation & public utilities wages with complex services	+0.50
Percentage of families earning less than $4,000 yearly with ratio of operations/discharges	-0.53

Exogenous Measures

Manufacturing wages with percentage of families earning less than $4,000 yearly	-0.69
Manufacturing wages with family income	+0.73
Transportation & public utilities wages with Percentage of families earning less than $4,000	-0.61
Transportation & public utilities wages with family income	+0.69
Transportation & public utilities wages with Ob/Gyn's per 10,000 population	+0.53
Retail wages with percentage of families earning less than $4,000 yearly	-0.56
Retail wages with Ob/Gyn's per 10,000 population	+0.65
Retail wages with direct care specialists per 10,000 population	+0.52
Retail wages with other specialists per 10,000 population	+0.53
Retail wages with surgical specialists per 10,000 population	+0.50
Hospital wages with family income	+0.56

HCFA Measures

Family income with total number of hospital beds	0.42
Family income with percentage within an SMSA	0.66
Total number of hospital beds with percentage within an SMSA	0.56

The factors found in this case appeared to be well defined; each measure tended to load highly on only one factor. Therefore, it was relatively straightforward to interpret each factor. The first factor appeared to be the input price factor, the third factor appeared to be a case mix severity factor, and the other ractors appeared to measure case mix.

Further interpretation is provided below.[47]

Factor 1: The four input price measures, hospital wages (.59), manufacturing wages (.63), transportation and public utilities hourly wages (.77) and retail hourly wages (.74), have their highest loadings on this factor. One measure--the ratio of surgical operations to discharges (.48)--which was hypothesized to represent case mix, also loads highest on this factor, although it loads highly on factors 2, 3, and 4 (.38, .25, and .25) as well (thereby providing considerable input to the two case mix factors). Another measure--the percentage of families earning less than $4000--also loaded highest on this factor (-.60). Its loading on the third factor, which apparently was the case mix severity, factor was -.56.

Factor 2: The three service measures--the number of quality enhancing services (.55), the number of complex services (.83), and the number of basic services (.68)--all load highest on this case mix factor.

Factor 3: The measure, high demand ages, has its highest loading (-.55) on this case mix severity factor; in addition, the other hypothesized case mix severity measure, percentage of families earning less than $4,000 annually, has its second highest loading (-.56) on this factor.

Factor 4: The measure of the number of basic services loads the highest on this factor (.49), and the number of quality enhancing services has its second highest loading (.53) on this case mix factor.

Factor 5: The ratio of outpatient visits to discharges has its highest loading (.50) and the number of community services has its second highest loading (.49) on this additional case mix factor.

Factor 6: The ratio of births to discharges has its only sizeable loading (.36) on this fourth and final case mix factor.

The set of exogenous measures was also subjected to factor analysis, using the same eigenvalue limit (1.0) and varimax rotation. In this case, five factors were found which accounted for 83.1 percent of the total variation. Three factors were apparently case mix factors (factors 1, 3, and 4), the second factor was apparently the input price factor, and one factor (the fifth) was difficult to interpret and therefore discarded due to the fact that only the measure of the labor force participation rate loaded highly on this factor. Individual factor interpretations are given below.

---

[47] The numbers in parentheses indicate factor loadings.

Factor 1: Every measure of the supply of medical personnel (i.e., OB/Gyn's per 10,000 population (.81), primary care M.D.'s per 10,000 population (.95), medical specialists per 10,000 population (.95), specialists per 10,000 population (.95), and surgical specialists per 10,000 population (.94)) loaded very highly on this factor, clearly identifying it as a case mix factor.

Factor 2: The four input price measures load highest on this factor, with the three county based wage rates having a minimum loading of 0.74 and the hospital based wage rate having a loading of 0.57. In addition, the percentage of families earning less than $4000 annually (-.80) as well as family income (.87) load highest on this factor. The measures for the labor force participation rate and the disabled rate have their second highest loadings on this factor (.53 and -.39, respectively), supporting the assumption that this factor is extracting the "prices" variance in all measures.

Factor 3: One measure has its highest loadings on this factor--the percentage of non-whites (.73), and another measure--the disabled rate--has its second highest loading (.35) on this additional case mix factor.

Factor 4: This factor defines the third case mix factor. The measure of high demand ages (including the percentage of females aged 15 to 44) (.45), the disabled rate (.42), and the percentage of M.D.'s older than 60 (.27) have their highest loadings on this factor.

In addition to the two sets of case mix and input price measures, the measure for the rural markets variable (i.e., the uniform pressure occupancy index) was included in a preliminary factor analysis. Not surprisingly, it was found that this measure always loaded highest on the input price factor, since the variation in rural-urban differences is highly associated with price differences. However, since the uniform pressure occupancy index is the only measure for rural markets, it was not necessary to include this measure in the subsequent factor analyses. The standardized values of the uniform pressure occupancy index, comparable to the standardized factor scores, were therefore treated as an additional score to be included in the computation of distance measures. On balance, the conceptual advantage of having a representative measure for rural markets was believed to outweigh the effects of the implicit weighting resulting from any colinearity between the uniform pressure occupancy index and input prices.

Factor analysis was also performed on the variables used by HCFA in their cross classification system. Two factors were found which accounted for 100 percent of the total variation. The family income measure loaded highest on one factor, and the number of total beds loaded highest on the second factor. The dichotomous variable indicating the location of a hospital within a SMSA loaded approximately equally on both factors (0.65 and 0.61, respectively).

C. Variable Weight Selection

The use of factor scores in place of the actual variable measures removed any implicit weightings due to multicollinearities. Two explicit weighting schemes were then used to define the relative importance of each variable.

The first weighting scheme attempted to equalize the weights among each hypothesized variable (i.e., input price variable, the rural markets variable, etc.) under the assumption that there is no á priori reason to consider one variable more important than any other in the determination of hospital costs. Given the use of factor scores, however, and the fact that more than one factor had been identified for some variables (e.g. case mix), it was necessary to reduce the weight of some factors below unity. If, for example, a variable had several clearly associated factors, each associated factor was assigned an equal proportion of the unit weight. The factor weights determined in this manner are displayed in Tables 4.2, 4.3, and 4.4.

The second weighting scheme relaxed the assumption that all variables contribute equally to hospital costs, and attempted to determine each variable's relative importance by regressing the factor scores and the uniform pressure occupancy index on cost per case.[48] If a factor was statistically significant in the regression equation, the standardized regression coefficient for that factor was used as its weight. A weight of zero was assigned to factors whose coefficients were not significantly different from zero. The regression weights found for the three sets of variables are shown in Table 4.5. It is easily seen that these weights differ significantly from those developed by the unit weighting scheme.[49]

The use of the two weighting systems and the three sets of measures produced a total of six data sets to be analyzed. These six sets and their respective mnemonics are as follows:

1) Endogenous variables with unit weights (EnU)
2) Exogenous variables with unit weights (ExU)
3) HCFA variables with unit weights (HCFAU)
4) Endogenous variables with regression weights (EnR)
5) Exogenous variables with regression weights (ExR)
6) HCFA variables with regression weights (HCFAR).

---

[48]The use of cost per case is an admittedly inadequate and gross representation of cost, but it was used as the best alternative of the data available.

[49]Numerous other explicit weighting schemes exist. For example, weights for any factor might have been determined by the percentage of explained variance accounted for by that factor in the factor analysis. Other possibilities include the use of the Automatic Interaction Detector (AID) (Morgan and Sonquist, 1973) to eliminate the dependency on a predefined functional form, or an iterative scheme which re-estimates the weights as each group of homogeneous hospitals is found. In future work, a number of these approaches will be explored.

TABLE 4.2

Factor Analysis: Endogenous Variables - N = 1070 Hospitals

VARIABLE		MEASURE*	UNIT WEIGHT DETERMINATION
Input Prices	Factor 1:	Manufacturing Hourly Wages Public & Transportation Hourly Wages Retail Hourly Wages Hospital Yearly Wages % Families Earning Less Than $4,000 Ratio of Surgical Operations to Discharges	1.0
Case Mix	Factor 2:	Number of Quality Enhancing Services Number of Complex Services Number of Community Services (Ratio of Surgical Operations to Discharges)	.25
	Factor 4:	Number of Basic Services (Number of Quality Enhancing Services)	.25
	Factor 5:	Ratio of Outpatient Visits to Discharges (Number of Community Services)	.25
	Factor 6:	Ratio of Births to Discharges	.25
Case Mix Severity	Factor 3:	High Demand Ages (% Families Earning Less Than $4,000)	1.0
Rural Markets	Standardized Values of Uniform Pressure Occupancy Index		1.0

* Measures in parentheses have a sizeable, but not maximum, loading on this factor.

TABLE 4.3

Factor Analysis: Exogenous Variables - N = 1070 Hospitals

VARIABLE		MEASURE*	UNIT WEIGHT DETERMINATION
Input Prices	Factor 2:	Manufacturing Hourly Wages Public & Transportation Hourly Wages Retail Hourly Wages Hospital Wages Family Income % Families Earning Less Than $4,000 Labor Force Participation Rate (Disabled Rate)	1.0
Case Mix	Factor 1:	OB/Gyn's per 10,000 Population Primary Care M.D.'s per 10,000 population Medical Specialists per 10,000 Population Other Direct Care Specialists per 10,000 Pop. Other Specialists per 10,000 Population Surgical Specialists per 10,000 Population	.33
	Factor 3:	% Non-White (Disabled Rate)	.33
	Factor 4:	High Demand Ages (With % Females 15-44) Disabled Rate % M.D.'s Aged Over 60	.33
Rural Markets		Standardized Values of Uniform Pressure Occupancy Index	1.0

* Measures in parentheses have a sizeable, but not maximum, loading on this factor.

TABLE 4.4

Factor Analysis: HCFA variables - N = 1070 Hospitals

VARIABLE		MEASURE*	UNIT WEIGHT DETERMINATION
Income	Factor 1:	Median Family Income SMSA/non-SMSA	1.0
Bed Size	Factor 2:	Total Bed (SMSA/non-SMSA)	1.0

*Measures in parentheses have a sizeable, but not maximum, loading on this factor.

TABLE 4.5

Regression Weight Determination* - N = 1053 Hospitals

VARIABLES	FACTOR	REGRESSION COEFFICIENT	STANDARDIZED REGRESSION COEFFICIENT**	SIGNIFICANCE †† OF FACTOR	R^2 CHANGE
ENDOGENOUS	Input Prices (Factor 1)	161.68	.267	.000	.1736
	Case Mix Severity (Factor 3)	-29.03	-.039	.118 †	.0000
	Case Mix (Factor 2)	149.10	.251	.000	.0889
	Case Mix (Factor 4)	- 2.23	-.003	.908 †	.0000
	Case Mix (Factor 5)	333.12	.446	.000	.2243
	Case Mix (Factor 6)	-69.37	-.066	.007	.0025
	Uniform Pressure Occupancy Index	77.54	.145	.000	.0064
	(Constant)	850.73			
	TOTAL			.000	.4957
EXOGENOUS	Input Prices (Factor 2)	181.87	.330	.000	.1571
	Case Mix (Factor 1)	198.66	.375	.000	.1672
	Case Mix (Factor 3)	-21.86	-.035	.231 †	.0000
	Case Mix (Factor 4)	31.11	.054	.035	.0041
	Uniform Pressure Occupancy Index	61.12	.115	.007	.0047
	(Constant)	852.52			
	TOTAL			.007	.3329
HCFA	Factor 1 (Income)	95.82	.134	.000	.1533
	Factor 2 (Bed Size)	295.21	.380	.000	.0782
	(Constant)	851.03			
	TOTAL			.000	.2316

* Dependent variable used was cost per case.defined as total expenses divided by total discharges.

** The standardized regression coefficients are the weights applied to the factors.

† These factors were given weights of 0 since they are not significant at the .05 level.

†† The significance level is based on the F ratio.

II. Cluster Analysis Results: Subsample of 194 Hospitals

In order to test the feasibility of the methodology and to test sample versus population results, a random subsample of 194 hospitals was selected from the complete data base of 1070 hospitals. On the basis of the sampling procedure, as well as subsequent examination of the subsample, it appeared that the subsample accurately mirrored the complete sample.

Based on the three sets of measures and two sets of weights, six sets of similarity measures were calculated. In each set, the similarity measure between two hospitals was the squared Euclidean distance calculated from the weighted factor scores. As described in the third chapter, six clustering algorithms (complete linkage, average linkage between groups, average linkage within groups, median method of Gower, centroid method, and Ward's suboptimization method) were then applied to each similarity matrix. The outputs from these algorithms were combined to produce a single composite dendrogram for each of the six variable sets. These composite dendrograms were then analyzed to find reasonable sets of hospital partitions.

A composite similarity measure between two hospitals represents the (weighted) average joining class level at which those two hospitals were joined by the six individual algorithms. In order to compute such an average, each algorithm's results are weighted by the square of its respective absolute cophenetic correlation coefficient (i.e., the correlation coefficient between the vector of pairwise similarity measures and the vector of joining class levels). The absolute cophenetic correlation coefficients for the six variable sets and the six individual algorithms are shown in Table 4.6, and empirically support the use of a composite dendrogram. An examination of Table 4.6 reveals that no single individual method consistently outperforms any other method (the maximum cophenetic correlation coefficient for each data set is indicated by a "*"). Of the six methods used, three of the methods have maximum values at some time. On the other hand, it is readily apparent that in most cases the differences between methods, at least in terms of the cophenetic correlation coefficient, are not significantly different. On the basis of the results in Table 4.6, it would be difficult to justify selecting the results from one algorithm with a correlation coefficient of, say, 0.74 while completely neglecting the results from another algorithm with a correlation coefficient of 0.73 (which would be the case for the EnU approach).

Given the six weighted similarity matrices, the composite dendrograms were found for all six data sets: EnR, ExR, HCFAR, EnU, ExU, and HCFAU. These composite dendrograms and their respective group structures are presented in Figures 4.1 to 4.6.

Each composite dendrogram was analyzed using the concept of expected distinctiveness defined in the third chapter. The results of these analyses, given in Table 4.7 (for unit weights) and Table 4.8 (for regression weights), summarize the optimal and suboptimal group structures as determined by the values of expected distinctiveness. The number following the partition identification (i.e., optimal or suboptimal) indicates the number of hospital groups in that partition; the precise definition of these groups is given under the "group structure"

TABLE 4.6

Absolute Cophenetic Correlation Coefficients For Clustering Algorithms and Variable Sets**
N = 194 Hospitals

	COMPLETE LINKAGE	AVERAGE BETWEEN GROUPS	AVERAGE WITHIN GROUPS	MEDIAN	CENTROID	WARDS
ENDOGENOUS UNIT (EnU) WEIGHTS	.665	.741*	.707	.728	.724	.680
EXOGENOUS UNIT (ExU) WEIGHTS	.722	.778*	.714	.763	.774	.674
HCFA UNIT (HCFAU) WEIGHTS	.884	.893	.894*	.887	.893	.890
ENDOGENOUS REGRESSION (EnR) WEIGHTS	.555	.731*	.633	.606	.716	.627
EXOGENOUS REGRESSION (ExR) WEIGHTS	.715	.727*	.710	.695	.699	.686
HCFA REGRESSION (HCFAR) WEIGHTS	.893	.900	.894	.897	.901*	.891

* indicates maximum value
** Note: All values are significant at the .001 level.

TABLE 4.7

Partition Summary: Unit Weights - N = 194 Hospitals

VARIABLE SET	PARTITION	% CORRECTLY CLASSIFIED*	E(D)**	GROUP STRUCTURE***
ENDOGENOUS	Optimal - 9	92.3	1133	(112-9-4-6-4-27-7-24-1)
	Subopt 1 - 12	91.8	1130	(112-3-1-4-1-4-6-4-27-7-24-1)
	Subopt 2 - 10	93.3	1129	(112-9-4-6-4-27-7-2-22-1)
	Subopt 3 - 13	93.3	1126	(112-3-1-4-1-4-7-4-27-7-2-22-1)
	Subopt 4 - 14	93.8	901	(73-3-1-2-4-29-9-4-6-4-27-7-24-1)
	Subopt 5 - 22	91.8	810	(73-3-1-2-4-29-3-1-4-1-4-6-4-2-1-1-3-20-7-2-22-1)
EXOGENOUS	Optimal - 5	95.4	1380	(112-1-45-34-2)
	Subopt 1 - 6	95.9	1270	(112-1-14-31-34-2)
	Subopt 2 - 8	90.7	1219	(46-9-31-26-1-45-34-2)
	Subopt 3 - 9	90.7	1109	(46-9-31-26-1-14-31-34-2)
	Subopt 4 - 12	89.7	898	(46-9-31-26-1-14-31-12-3-1-18-2)
HCFA	Optimal - 3	100.0	1035+M †	(27-87-80)
	Subopt 1 - 4	100.0	915+M	(27-10-77-80)
	Subopt 2 - 6	99.5	838+M	(27-10-17-45-15-80)
	Subopt 3 - 5	97.9	1558	(27-10-77-32-48)

* Percent of cases correctly classified by linear discriminant analysis.

** Expected distinctiveness.

*** Groups as identified sequentially from top to bottom of respective dendrogram.

† M represents a very large real number (M>> 0)

TABLE 4.8

Partition Summary: Regression Weights - N 194 Hospitals

APPROACH	PARTITION	% CORRECTLY CLASSIFIED*	E(D)**	GROUP STRUCTURE***
ENDOGENOUS	Optimal - 7	97.9	1146	(164-2-2-3-9-1-13)
	Subopt 1 - 11	95.4	1108	(100-54-3-6-1-2-2-3-9-1-13)
	Subopt 2 - 8	94.8	1103	(164-2-2-3-9-1-11-2)
	Subopt 3 - 27	87.6	915	(19-3-2-4-9-1-27-5-3-1-1-1-1-3-1-6-13-54-3-6-1-2-2-3-9-1-13)
EXOGENOUS	Optimal - 5	94.8	1781	(91-56-44-1-2)
	Subopt 1 - 9	93.3	1774	(91-56-3-6-2-4-29-1-2)
	Subopt 2-- 10	94.8	1385	(91-4-4-4-11-10-23-44-1-2)
	Subopt 3 - 14	90.2	1378	(91-4-4-4-11-10-23-3-6-2-4-29-1-2)
	Subopt 4 - 19	95.4	793	(19-7-21-13-11-20-4-4-4-11-10-23-3-6-2-4-29-1-2)
HCFA	Optimal - 3	100.0	1156+M†	(104-10-80)
	Subopt 1 - 5	98.5	886+M	(34-44-26-10-80)

* Percent of cases correctly classified by linear discriminant analysis.

** Expected distinctiveness.

*** Groups as identified sequentially from top to bottom of respective dendrogram.

† M represents a very large real number (M>>0)

Composite Dendrogram: Endogeneous Approach, Regression Weights

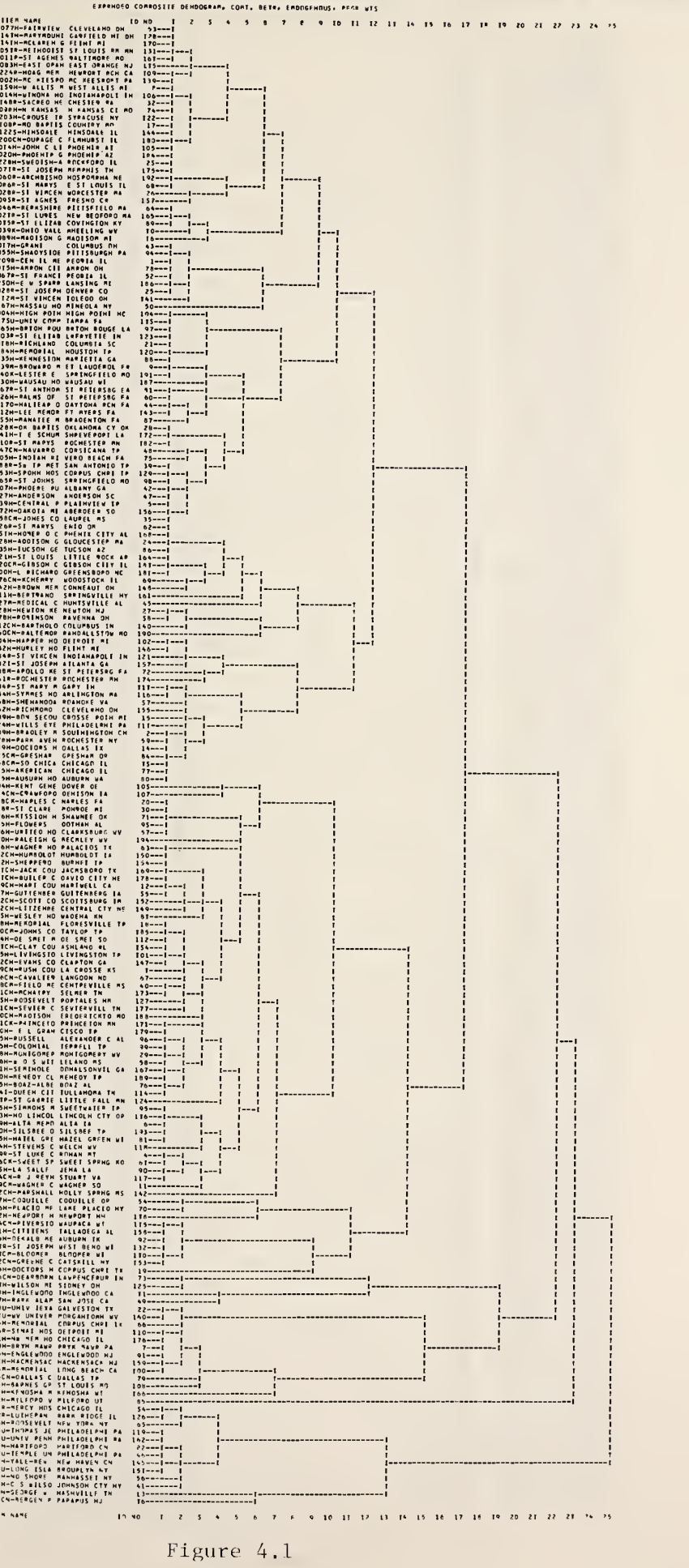


Figure 4.1

Composite Dendrogram: Exogenous Approach, Regression Weights

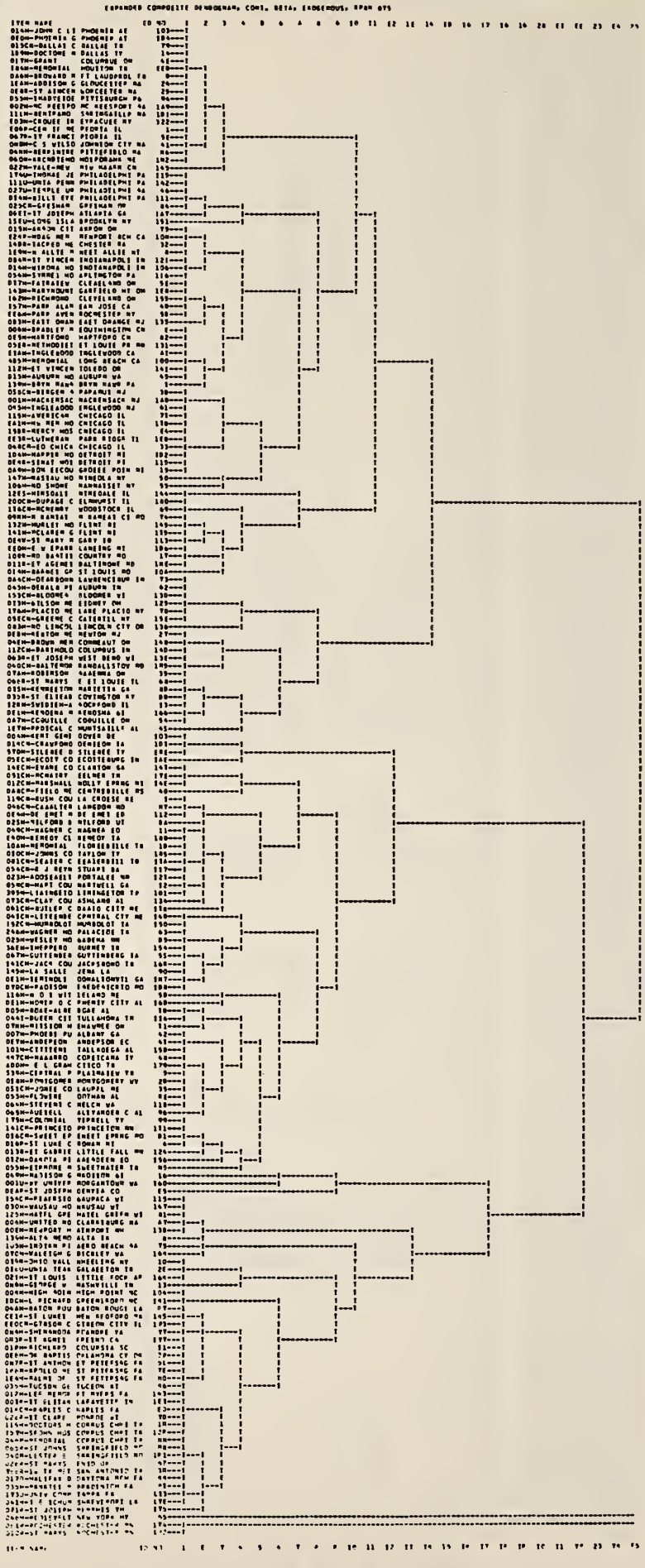


Figure 4.2

Composite Dendrogram: HCFA Approach, Regression Weights

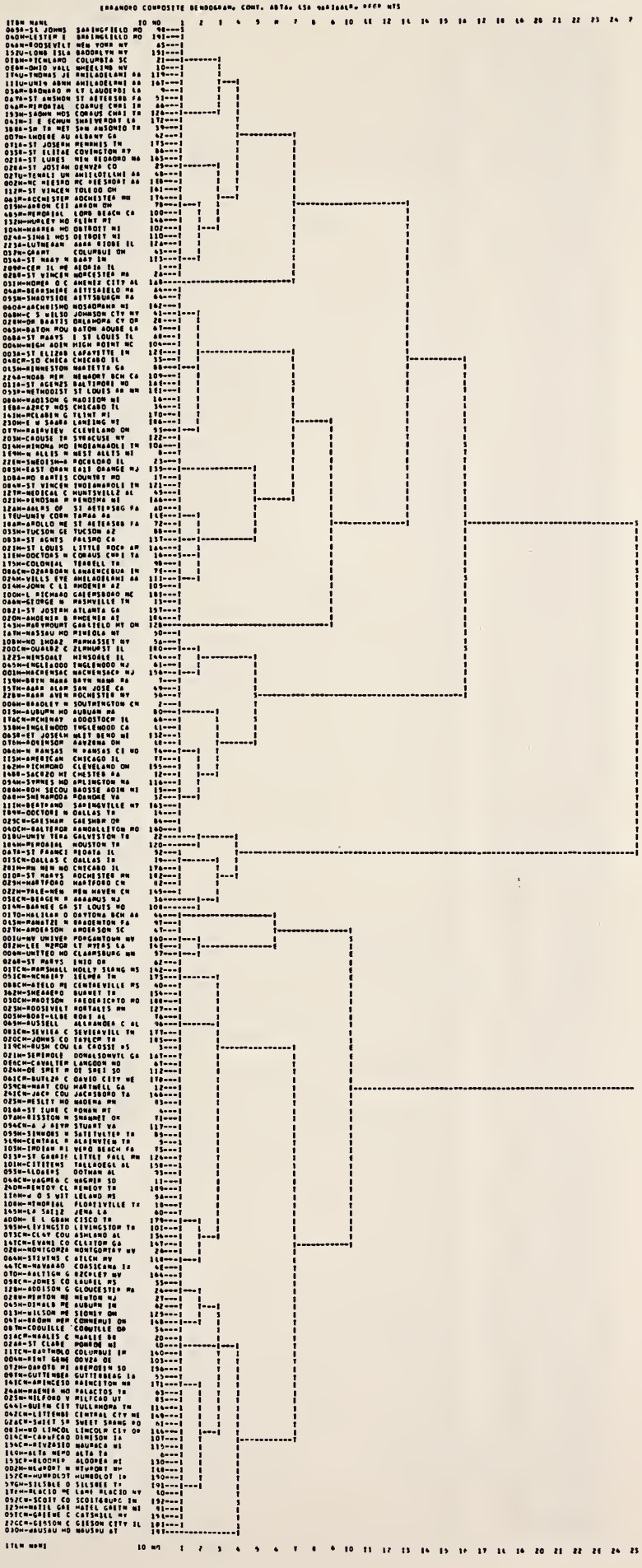


Figure 4.3

Composite Dendrogram: Endogenous, Unit Weights

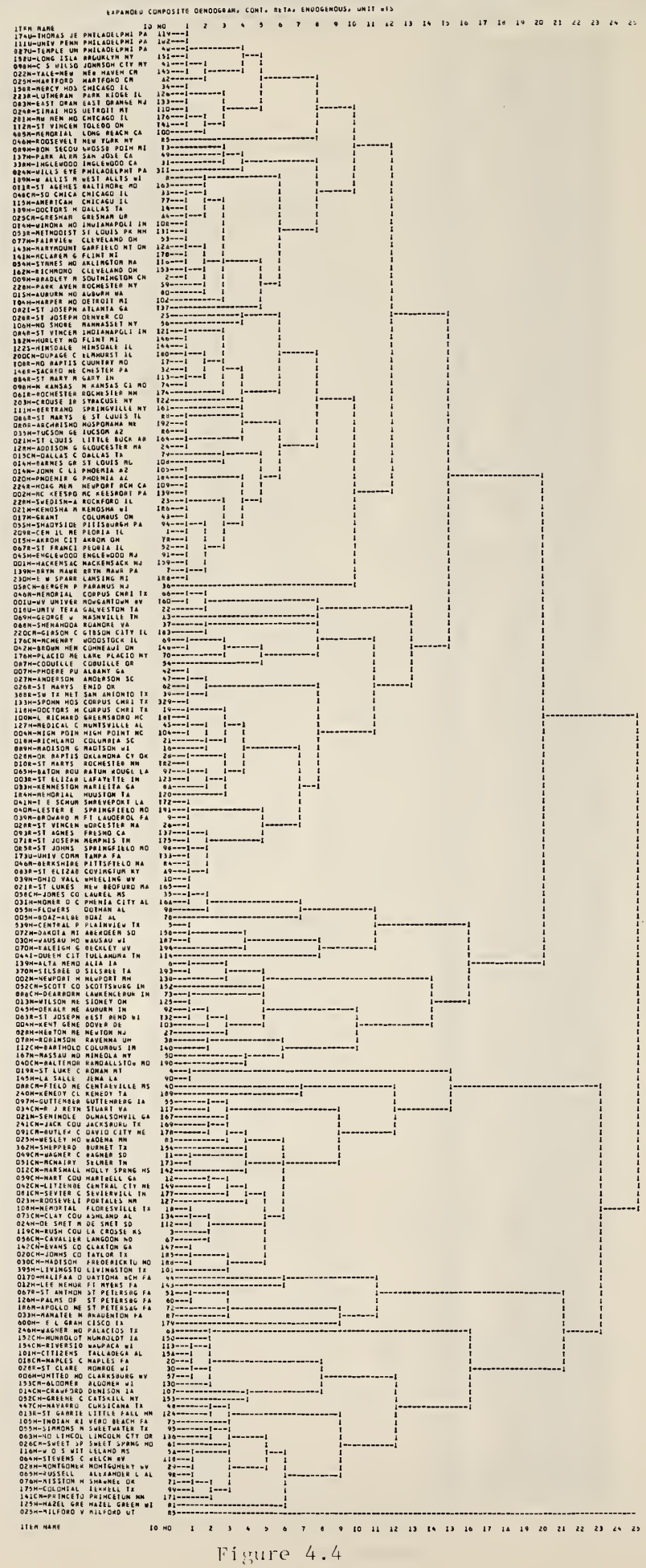


Figure 4.4

Composite Dendrogram: Exogenous Approach, Unit Weights

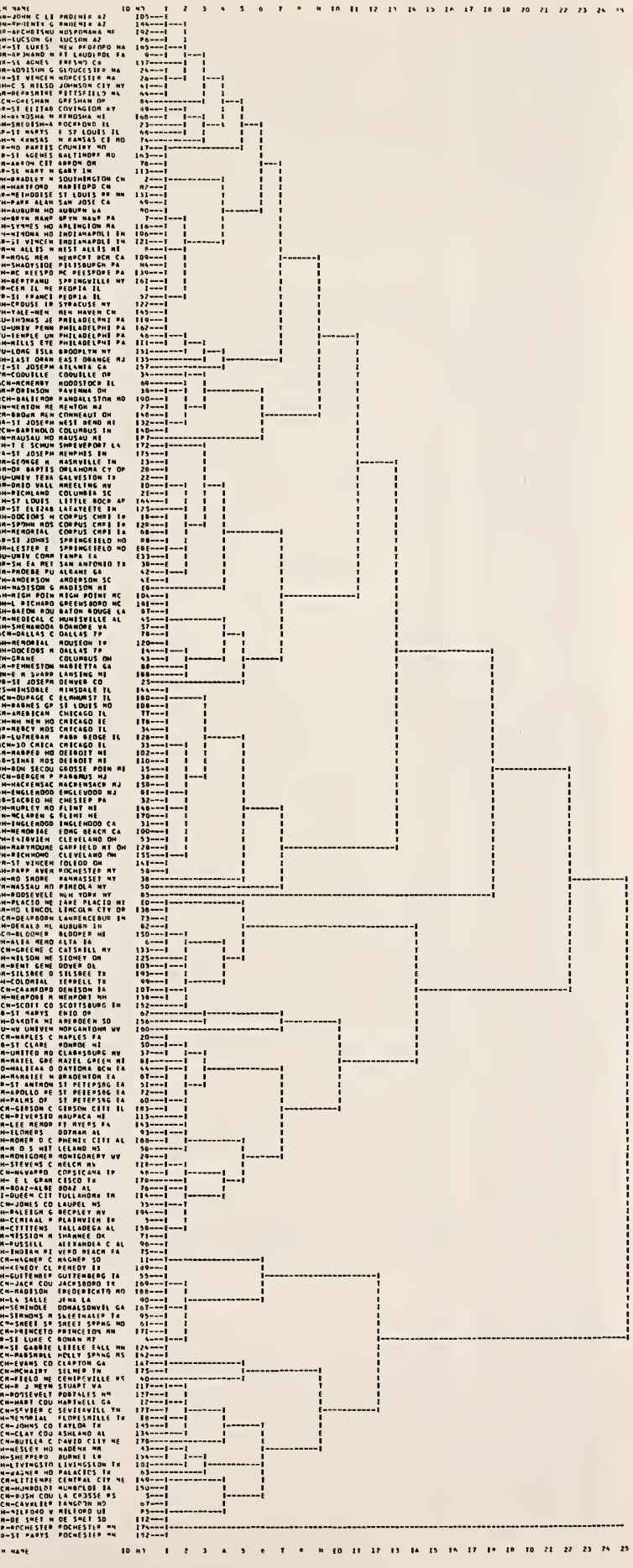


Figure 4.5

Composite Dendrogram: HCFA Approach, Unit Weights

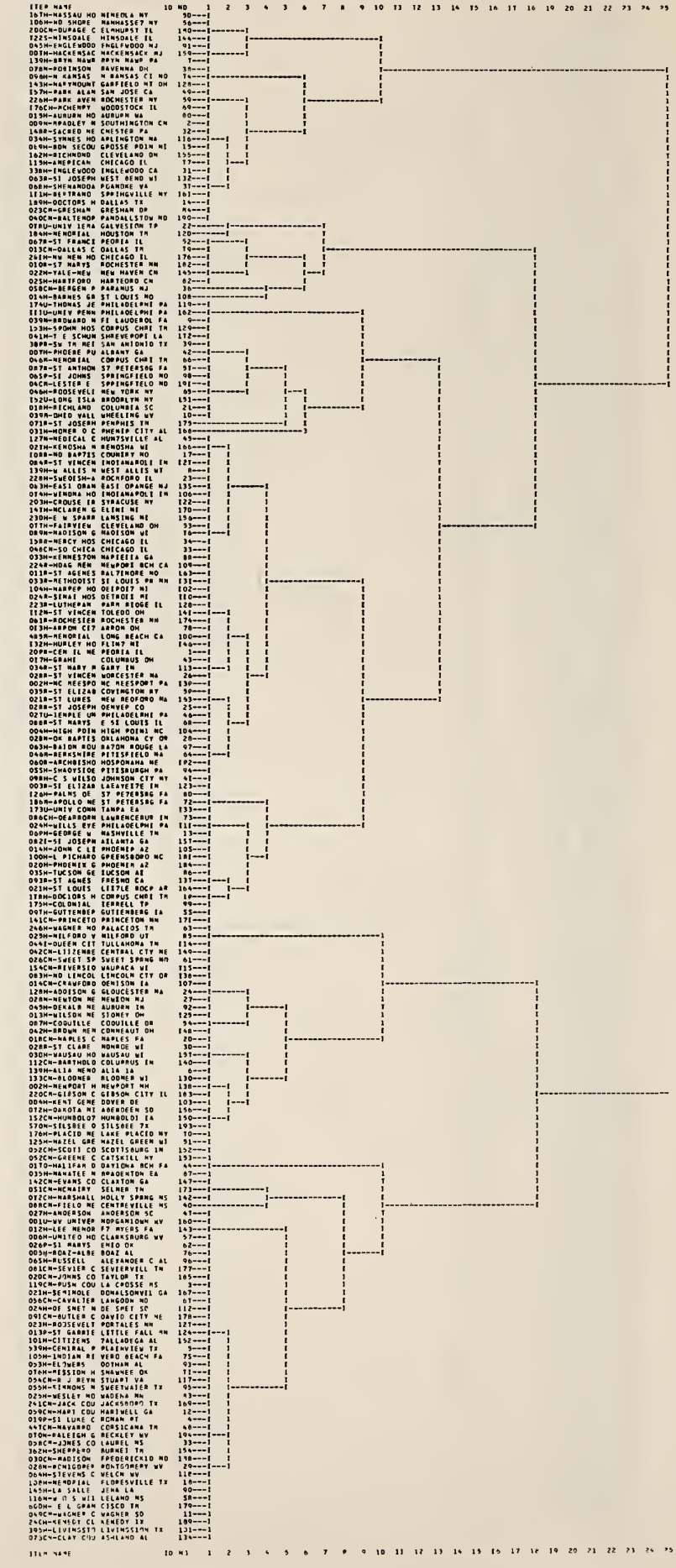


Figure 4.6

heading. For example, in the EnU variable set, the first 112 hospitals in that composite dendrogram are grouped together, the next 9 hospitals form the second group, etc. in the optimal partition. It is interesting to note that the percentage of hospitals correctly classified by a linear discriminant analysis closely corresponds to the values of the expected distinctiveness measure. From these tables, it would appear that the partitions can be well discriminated by linear functions, and that the values of expected distinctiveness reflect this ability.

Further partition validation was provided by several one-way analyses of variance (Table 4.9) using the identified groups as independent variables, and the weighted factors used to create the original similarity measures as dependent variables.[50] In the case of unit weights, all factors were significant at the .05 level; in fact, all factors except one case mix factor (factor 5: weight=0.25) were significant at the .01 level. When using regression weights, however, the results were significantly different. In this case factors 3 and 4 in the endogenous variable set are not significant in determining hospital groups; examination of Table 4.5 reveals that the weights on both these factors were zero since they were not statistically significant in the regression equation from which the weights were drawn.

III. Cluster Analysis Results: 1070 Hospital Sample

Three variable sets were used in the analysis of the full sample: EnR, EnU, and ExR. Due to cost limitations, it was necessary to restrict the number of algorithms used for each variable set. Therefore, a composite dendrogram based on three representative algorithms (instead of the six used to analyze the 194 hospital subsample) was produced for the endogenous variable set with regression weights, and a single algorithm was used for the other two variable sets.

The selection of algorithms for the EnR approach was determined by the size of the absolute cophenetic correlation coefficients between clustering algorithms and their respective similarity matrices (Table 4.6), and by type of algorithm. An attempt was made to group the algorithms themselves on the basis of mathematical similarity, and the algorithm with the highest cophenetic correlation coefficient was then chosen for each group. This admittedly inexact selection process resulted in the use of the complete linkage, average linkage between groups, and centroid methods. On the basis of their performance in analyzing the 194 hospital subsample, it was felt that these algorithms were most likely to result in a composite dendrogram most similar to the composite dendrogram which would have been determined if all six algorithms had been used.

For the EnU and ExR approaches, the average linkage between groups algorithm was selected since it had the highest absolute cophenetic correlation coefficient among all approaches for the hospital subsample (Table 4.6).

---

[50] The individual factors and their respective unit weights are identified in Tables 4.2, 4.3, and 4.4; the regression weights used for each factor are given in Table 4.5. The last factors for the endogenous and exogenous variable sets in Table 4.9 (i.e., factors 7 and 5, respectively) represent the uniform pressure occupancy index.

TABLE 4.9

Analysis of Variance - N = 194 Hospitals

	UNIT WEIGHTS						REGRESSION WEIGHTS					
	BETWEEN GROUP		WITHIN GROUP				BETWEEN GROUP		WITHIN GROUP			
FACTOR	D.F.	MEAN SQ	D.F.	MEAN SQ	F RATIO	F PROB	D.F.	MEAN SQ	D.F.	MEAN SQ	F RATIO	F PROB
					ENDOGENOUS VARIABLE SET							
1	8	11.200	185	.358	31.230	.000	6	5.230	187	.666	7.852	.000
2	8	5.109	185	.544	9.389	.000	6	8.936	187	.470	19.003	.000
3	8	9.045	185	.200	45.020	.000	6	.627	187	.565	1.110	.358
4	8	1.983	185	.281	7.045	.000	6	.298	187	.353	.843	.538
5	8	.756	185	.383	1.974	.052	6	7.837	187	.160	48.988	.000
6	8	.848	185	.187	4.533	.000	6	.907	187	.192	4.719	.000
7	8	20.691	185	.136	151.179	.000	6	5.022	187	.859	5.843	.000
					EXOGENOUS VARIABLE SET							
1	4	62.909	189	.199	315.052	.000	4	63.706	189	.182	348.465	.000
2	4	32.411	189	.412	78.488	.000	4	41.997	189	.210	199.927	.000
3	4	11.219	189	.536	20.905	.000	4	3.602	189	.697	5.162	.001
4	4	4.209	189	.770	6.125	.000	4	5.697	189	.670	8.494	.000
5	4	38.227	189	.200	190.409	.000	4	28.963	189	.396	72.988	.000
					HCFA VARIABLE SET							
1	3	34.099	190	.088	387.564	.000	2	47.390	191	.126	373.470	.000
2	3	28.562	190	.041	685.760	.000	2	40.762	191	.063	644.727	.000

The determination and examination of suboptimal partitions proceed in the same manner used to analyze the smaller subsample as described in the third chapter. Examination of suboptimal partitions was discontinued when it became evident that further breakdown would result in a very large decrease in the overall expected distinctiveness value, E(D), and the atomization of the partition into many small groups. Tables 4.10, 4.11, and 4.12 provide information on the configuration of the optimal and suboptimal partitions found for the three approaches examined. Descriptions of the groups found in this analysis are given in Appendix A.

Cluster validation proceeded in a fashion similar to the approach used for the smaller subsample. Linear discriminant analysis was initially performed; the results are presented in Table 4.13. As with the subsample results, there appears to exist a close relationship between the value of expected distinctiveness and the percentage of hospitals correctly classified by the linear discriminant functions. Moreover, the percentage of hospitals correctly classified is strikingly high, which again strongly suggests that the hospital groups are fairly well-defined, compact, a separable by linear functions. It also suggests that once a sample of hospitals has been clustered and used to determine the discriminant functions, all hospitals might be classified by these discriminant functions with a fairly high degree of accuracy.

The analysis of variance results (Tables 4.14, 4.15, and 4.16) are informative in several ways. First, all factors in every group structure have highly significant F ratios (significant at a .001 level), indicating that every group structure successfully discriminates on the factors used to create the structures.

It is also interesting to examine the absolute values of the F ratios, which give the ratio of the between group sum of squares to the within group sum of squares (i.e., a high F ratio indicates that most of a measure's variation occurs between groups rather than within groups). The results in Tables 4.14, 4.15 and 4.16 indicate that a strong relationship exists between the weights attached to each factor and their respective F values. In almost all cases, those factors with the highest weights also have the highest F ratios; for example, in the exogenous variable set with regression weights, factor 1 has the largest weight (.375) and greatest associated F ratio for all partitions.

In order to test the relationship between the factor weights and F ratios further, the simple product-moment correlation coefficient was calculated between these two vectors for all three approaches. The analysis of all three variable sets resulted in values of 0.77 or higher. In addition, given the small differences between factor weights and the significant differences between F values in some cases, it would also appear from this analysis that resultant clusters are fairly sensitive to changes in the similarity measures in general and the weights in particular.

A. Interpretation/Evaluation of Group Structures

The optimal and suboptimal (in terms of the expected distinctiveness) partitions found for the three variable sets (i.e., EnR, EnU, and ExR) were compared using the Λ measure developed in the third chapter. Basically, the Λ measure is a

TABLE 4.10

Partitions: Endogenous Approach - Regression Weights (N = 1070 Hospitals)

Optimal (E(D) = 4887)

Group	Size
1	737
2	174
3	53
4	1
5	1
6	20
7	3
8	1
9	3
10	3
11	69
12	4
13	1

13 Groups Total

Subopt 1 (E(D) = 3947)

(splits #1 of optimal)

Group	Size
101	312
102	32
103	1
104	1
105	30
106	48
107	4
108	17
109	13
110	100
111	32
112	14
113	12
114	122
115	8

27 Groups Total

Subopt 2 (E(D) = 3771)

(splits #101 of subopt 1)

Group	Size
10101	14
10102	50
10103	7
10104	12
10105	25
10106	30
10107	74
10108	19
10109	81

35 Groups Total

TABLE 4.11

Partitions: Endogenous Approach - Unit Weights (N = 1070 Hospitals)

Optimal E(D) = 4888)

Group	Size
1	788
2	42
3	1
4	12
5	1
6	2
7	1
8	4
9	216
10	1
11	1
12	1

12 Groups Total

Subopt 3 (E(D) = 2748)

(splits #1 of optimal)

Group	Size
101	400
102	6
103	13
104	7
105	2
106	216
107	36
108	33
109	25
110	35
111	15

26 Groups Total

Subopt 1 (E(D) = 4479)

(splits #9 of optimal)

Group	Size
901	214
902	1
903	1

14 Groups Total

Subopt 2 (E(D) = 4327)

(splits #101 of subopt 1)

Group	Size
90101	113
90102	62
90103	39

16 Groups Total

TABLE 4.12

Partitions: Exogenous Approach - Regression Weights (N-1070 Hospitals)

Optimal (E(D) = 9700)

Group	Size
1	783
2	219
3	1
4	44
5	1
6	14
7	4
8	1
9	1
10	2

10 Groups Total

Subopt 3 (E(D) = 2614)
(splits #1 of optimal)

Group	Size
101	478
102	304
103	1

17 Groups Total

Subopt 2 (E(D) = 8873)
(splits #2 of optimal)

Group	Size
201	13
202	24
203	182

15 Groups Total

Subopt 1 (E(D) = 9480)
(splits #4 of optimal)

Group	Size
401	19
402	15
403	5
404	5

13 Groups Total

TABLE 4.13

Linear Discriminant Analysis Results (N = 1070 Hospitals)

Variable Set	Partition	No. of Groups	E(D)*	% Correctly Classified
EnR	Optimal	13	4887	89.9
	Suboptimal - 1	27	3947	78.2
	Suboptimal - 2	35	3771	77.8
EnU	Optimal	12	4888	94.4
	Suboptimal - 1	14	4479	94.6
	Suboptimal - 2	16	4327	93.0
	Suboptimal - 3	26	2748	88.0
ExR	Optimal	17	9700	95.5
	Suboptimal - 1	13	9480	95.7
	Suboptimal - 2	15	8873	94.8
	Suboptimal - 3	17	2614	94.4

*Expected distinctiveness

TABLE 4.14

Analysis of Variance: Endogenous Approach - Regression Weights
(N = 1070 Hospitals)

OPTIMAL GROUP STRUCTURE* (13 Groups)	BETWEEN GROUP D.F.	BETWEEN GROUP MEAN SQ	WITHIN GROUP D.F.	WITHIN GROUP MEAN SQ	F RATIO	F PROB
Factor 1 (Prices : wt=.267)	12	31.77	1057	.43	74.27	.000
Factor 2 (Case Mix: wt=.251)	12	25.34	1057	.53	47.78	.000
Factor 3 (Severity: wt=.039)	12	13.20	1057	.37	35.40	.000
Factor 4 (Case Mix: wt=.003)	12	4.15	1057	.46	9.12	.000
Factor 5 (Case Mix: wt=.446)	12	26.59	1057	.22	123.58	.000
Factor 6 (Case Mix: wt=.066)	12	1.91	1057	.24	7.86	.000
Factor 7 (UPOI: wt=.145)	12	59.48	1057	.34	177.05	.000

1st SUBOPTIMAL STRUCTURE** (27 Groups)	BETWEEN GROUP D.F.	BETWEEN GROUP MEAN SQ	WITHIN GROUP D.F.	WITHIN GROUP MEAN SQ	F RATIO	F PROB
Factor 1 (Prices: wt=.267)	26	23.21	1043	.22	105.33	.000
Factor 2 (Case Mix: wt=.251)	26	24.70	1043	.21	115.84	.000
Factor 3 (Severity: wt=.039)	26	7.31	1043	.35	21.03	.000
Factor 4 (Case Mix: wt=.003	26	4.72	1043	.39	12.06	.000
Factor 5 (Case Mix: wt=.446)	26	16.52	1043	.11	147.40	.000
Factor 6 (Case Mix: wt=.066)	26	1.41	1043	.23	6.04	.000
Factor 7 (UPOI: wt=.145)	26	30.57	1043	.26	116.30	.000

2nd SUBOPTIMAL STRUCTURE*** (35 Groups)	BETWEEN GROUP D.F.	BETWEEN GROUP MEAN SQ	WITHIN GROUP D.F.	WITHIN GROUP MEAN SQ	F RATIO	F PROB
Factor 1 (Prices: wt=.267)	34	19.75	1035	.16	126.26	.000
Factor 2 (Case Mix: wt=.251)	34	20.92	1035	.15	141.09	.000
Factor 3 (Severity: wt=.039)	34	5.74	1035	.35	16.62	.000
Factor 4 (Case Mix: wt=.003)	34	3.94	1035	.38	10.28	.000
Factor 5 (Case Mix: wt=.446)	34	13.27	1035	.09	144.34	.000
Factor 6 (Case Mix: wt=.066)	34	1.67	1035	.23	5.03	.000
Factor 7 (UPOI: wt=.145)	34	24.66	1035	.22	110.66	.000

---

* Highest value of Expected Distinctiveness
** Second highest value of Expected Distinctiveness
*** Third highest value of Expected Distinctiveness

TABLE 4.15

Analysis of Variance: Endogenous Approach - Unit Weights (N = 1070 Hospitals)

OPTIMAL GROUP STRUCTURE (12 Groups)	BETWEEN GROUP D.F.	BETWEEN GROUP MEAN SQ	WITHIN GROUP D.F.	WITHIN GROUP MEAN SQ	F RATIO	F PROB
Factor 1 (Prices: wt=1.00)	11	31.12	1058	.46	67.06	.000
Factor 2 (Case Mix: wt= .25)	11	12.73	1058	.68	18.59	.000
Factor 3 (Severity: wt=1.00)	11	27.90	1058	.23	120.28	.000
Factor 4 (Case Mix: wt= .25)	11	6.80	1058	.43	15.78	.000
Factor 5 (Case Mix: wt= .25)	11	16.69	1058	.34	48.65	.000
Factor 6 (Case Mix: wt= .25)	11	2.42	1058	.24	10.10	.000
Factor 7 (UPOI: wt=1.00)	11	61.95	1058	.37	169.15	.000
1st SUBOPTIMAL STRUCTURE (14 Groups)	**BETWEEN GROUP D.F.**	**BETWEEN GROUP MEAN SQ**	**WITHIN GROUP D.F.**	**WITHIN GROUP MEAN SQ**	**F RATIO**	**F PROB**
Factor 1 (Prices: wt=1.00)	13	26.50	1056	.46	57.24	.000
Factor 2 (Case Mix: wt= .25)	13	11.25	1056	.68	16.50	.000
Factor 3 (Severity: wt=1.00)	13	23.74	1056	.23	102.85	.000
Factor 4 (Case Mix: wt= .25)	13	5.80	1056	.43	13.45	.000
Factor 5 (Case Mix: wt= .25)	13	14.26	1056	.34	41.69	.000
Factor 6 (Case Mix: wt= .25)	13	2.68	1056	.23	11.56	.000
Factor 7 (UPOI: wt=1.00)	13	52.47	1056	.37	143.24	.000
2nd SUBOPTIMAL STRUCTURE (16 Groups)	**BETWEEN GROUP D.F.**	**BETWEEN GROUP MEAN SQ**	**WITHIN GROUP D.F.**	**WITHIN GROUP MEAN SQ**	**F RATIO**	**F PROB**
Factor 1 (Prices: wt=1.00)	15	23.38	1054	.46	51.05	.000
Factor 2 (Case Mix: wt= .25)	15	10.21	1054	.68	15.12	.000
Factor 3 (Severity: wt=1.00)	15	23.30	1054	.19	121.05	.000
Factor 4 (Case Mix: wt= .25)	15	5.81	1054	.42	13.79	.000
Factor 5 (Case Mix: wt= .25)	15	12.54	1054	.34	36.86	.000
Factor 6 (Case Mix: wt= .25)	15	2.49	1054	.23	10.82	.000
Factor 7 (UPOI: wt=1.00)	15	51.67	1054	.28	185.27	.000
3rd SUBOPTIMAL STRUCTURE (26 Groups)	**BETWEEN GROUP D.F.**	**BETWEEN GROUP MEAN SQ**	**WITHIN GROUP D.F.**	**WITHIN GROUP MEAN SQ**	**F RATIO**	**F PROB**
Factor 1 (Prices: wt=1.00)	25	24.78	1044	.29	120.95	.000
Factor 2 (Case Mix: wt= .25)	25	14.78	1044	.47	31.16	.000
Factor 3 (Severity: wt=1.00)	25	14.87	1044	.17	85.88	.000
Factor 4 (Case Mix: wt= .25)	25	8.98	1044	.29	30.61	.000
Factor 5 (Case Mix: wt= .25)	25	8.16	1044	.33	24.86	.000
Factor 6 (Case Mix: wt= .25)	25	3.40	1044	.19	18.25	.000
Factor 7 (UPOI: wt=1.00)	25	27.80	1044	.12	318.14	.000

TABLE 4.16

Analysis of Variance: Exogenous Approach - Regression Weights (N = 1070 Hospitals)

OPTIMAL GROUP STRUCTURE (10 Groups)	BETWEEN GROUP D.F.	BETWEEN GROUP MEAN SQ	WITHIN GROUP D.F.	WITHIN GROUP MEAN SQ	F RATIO	F PROB
Factor 1 (Case Mix: wt=.375)	9	100.66	1060	.14	701.05	.000
Factor 2 (Prices: wt=.330)	9	57.13	1060	.43	133.19	.000
Factor 3 (Case Mix: wt=.035)	9	8.63	1060	.77	11.20	.000
Factor 4 (Case Mix: wt=.054)	9	7.61	1060	.47	16.18	.000
Factor 5 (UPOI: wt=.115)	9	67.41	1060	.44	154.57	.000

1st SUBOPTIMAL STRUCTURE (13 Groups)	BETWEEN GROUP D.F.	BETWEEN GROUP MEAN SQ	WITHIN GROUP D.F.	WITHIN GROUP MEAN SQ	F RATIO	F PROB
Factor 1 (Case Mix: wt=.375)	12	76.90	1057	.13	600.49	.000
Factor 2 (Prices: wt=.330)	12	43.93	1057	.42	105.16	.000
Factor 3 (Case Mix: wt=.035)	12	8.45	1057	.75	11.27	.000
Factor 4 (Case Mix: wt=.054)	12	7.07	1057	.46	15.50	.000
Factor 5 (UPOI: wt=.115)	12	50.67	1057	.44	116.18	.000

2nd SUBOPTIMAL STRUCTURE (15 Groups)	BETWEEN GROUP D.F.	BETWEEN GROUP MEAN SQ	WITHIN GROUP D.F.	WITHIN GROUP MEAN SQ	F RATIO	F PROB
Factor 1 (Case Mix: wt=.375)	14	66.02	1055	.13	520.43	.000
Factor 2 (Prices: wt=.330)	14	39.71	1055	.39	101.48	.000
Factor 3 (Case Mix: wt=.035)	14	13.72	1055	.66	20.61	.000
Factor 4 (Case Mix: wt=.054)	14	6.21	1055	.46	13.65	.000
Factor 5 (UPOI: wt=.115)	14	46.26	1055	.40	115.81	.000

3rd SUBOPTIMAL STRUCTURE (17 Groups)	BETWEEN GROUP D.F.	BETWEEN GROUP MEAN SQ	WITHIN GROUP D.F.	WITHIN GROUP MEAN SQ	F RATIO	F PROB
Factor 1 (Case Mix: wt=.375)	16	57.88	1053	.12	461.18	.000
Factor 2 (Prices: wt=.330)	16	50.06	1053	.16	313.94	.000
Factor 3 (Case Mix: st=.035)	16	13.89	1053	.64	21.76	.000
Factor 4 (Case Mix: wt=.054)	16	7.29	1053	.43	17.05	.000
Factor 5 (UPOI: wt=.115)	16	46.60	1053	.31	151.75	.000

weighted matching coefficient using all pairwise hospital combinations as observations, where the weights are determined under the hypothesis that grouping occurs at random. It was shown in the third chapter that random grouping results in an expected value of Λ equal to 0.5. Due to the large number of hospitals in this sample (i.e., 1070 hospitals) and the even larger number of pairwise combinations (i.e., 571,915 possible pairs), a 25 percent random sample of pairwise combinations was used in calculating Λ. Tests showed that little, if any, accuracy was lost by this approach.

The first set of comparisons examined the differences between partitions found with the EnR and the EnU approaches (i.e., the variables remained the same but the weights varied). Optimal partitions in both approaches resulted in 12 groups; the Λ measure calculated between these partitions showed them to be highly similar (Λ = 0.81). As evident from Table 4.17, the most distinctive partition for the EnR approach is highly similar with the most distinctive, first suboptimal, and second suboptimal partitions found for the EnU approach. Comparisons of the most distinctive EnR partition with the third suboptimal EnU partition, however, show little similarity; the null hypothesis that grouping has occurred randomly cannot be rejected with Λ equal to 0.462. Clues to the cause of this change in Λ are provided by the analysis of variance results. Using the F ratio as an indicator, the maximally distinctive EnR partition appeared to be heavily influenced by the UPOI measure (an urban-rural indicator) and case mix. The maximally distinctive, first suboptimal, and second suboptimal EnU partitions are also heavily influenced by the UPOI measure as well as the case mix severity factor. However, in the third suboptimal EnU partition, the influence of the case mix severity factor appears to decline. This partition apparently is determined mostly by an urban-rural indicator and input prices.

With one exception, the remaining comparisons remained significantly below 0.500, supporting the conjecture that once the EnR partitions begin to represent a more complex weight scheme, the two partition structures no longer correspond. Comparisons between the EnR and ExR approaches (Table 4.18), and the EnU and ExR approaches (Table 4.19) show similar patterns; the most distinctive partitions appear to be significantly related, but as the number of groups increase in the suboptimal partitions, the partition comparison measure decreases.

While some explanation may be gathered by examining the factors and the analyses of variance, this phenomenon in general (i.e., a lower partition comparison measure between partitions with larger number of groups) may be explained by examining the nature of the problem in general. The comparison between the optimal ExR partition (10 groups) and the third suboptimal EnU partition (26 groups) in Table 4.19 provides an example. The low Λ measure (0.368) would indicate that for the two partitions, it may be possible to combine the groups in the EnU partition into 10 groups in such a way that the resultant partition would be identical to the optimal ExR partition. In other words, while these two partitions appear to be very dissimilar, it may be that the partitions are quite consistent in the sense that one partition maps onto the other. This problem of finding the partition mapping which maximizes Λ will be studied in future work.

TABLE 4.17:

Partition Comparison: EnU Versus EnR* (N = 1070 Hospitals)

		EnR Approach		
		Optimal (12 groups)	1st Suboptimal (27 groups)	2nd Suboptimal (35 groups)
EnU Approach	Optimal (12 groups)	0.811	0.286	0.170
	1st Suboptimal (14 groups)	0.811	0.278	0.160
	2nd Suboptimal (16 groups)	0.844	0.281	0.161
	3rd Suboptimal (26 groups)	0.462	0.611	0.334

* Λ Partition Comparison Measure

TABLE 4.18

Partition Comparison: EnR Versus ExR*

ExR Approach	EnR Approach: Optimal (13 groups)	EnR Approach: 1st Suboptimal (27 groups)	EnR Approach: 2nd Suboptimal (35 groups)
Optimal (10 groups)	0.662	0.294	0.197
1st Suboptimal (13 groups)	0.666	0.279	0.174
2nd Suboptimal (15 groups)	0.655	0.261	0.163
3rd Suboptimal (17 groups)	0.462	0.321	0.269

* Λ Partition Comparison Measure

TABLE 4.19:

Partition Comparison: EnU Versus ExR*

ExR Approach	EnU Approach: Optimal (12 groups)	1st Subopt (14 groups)	2nd Subopt (16 groups)	3rd Subopt (26 groups)
Optimal (10 groups)	0.742	0.740	0.720	0.368
1st Subopt (13 groups)	0.745	0.735	0.719	0.357
2nd Subopt (15 groups)	0.723	0.732	0.712	0.355
3rd Subopt (17 groups)	0.486	0.484	0.475	0.489

*Λ Partition Comparison Measure

CHAPTER FIVE

SUMMARY AND CONCLUSIONS

The purpose of this study has been to provide an initial investigation into the feasibility of a hospital classification methodology that could be used as part of a prospective payment system in the U. S. hospital industry. The aim of this chapter is to briefly summarize the major conclusions suggested by our work, to reiterate the qualifications that must accompany these conclusions, and to suggest directions for further investigation. It is important to note that the primary focus of this initial effort was on the development of the conceptual and statistical methodology; while we have reported the results of an empirical analysis that is based on our proposed methodology, severe data limitations restrict the usefulness of the empirical results to answering particular questions about the characteristics of the system and suggesting potential problem areas for further study. The final groupings reported in Chapter 4, therefore, must be regarded as illustrative. They are not intended to reflect our view of the most appropriate industry classification for rate determination or rate review purposes.

I. Conceptual Framework

Perhaps the most important conclusion that can be drawn from the conceptual development in Chapter 2 is that the selection of the criteria on which hospital similarity is to be based is exceedingly important. While most of the attention in the literature has focused on the selection of the statistical technique to be used (e.g. factor analysis, cross classification, cluster analysis etc.), the use of inappropriate classification variables will lead to problems and biases equally as severe as those resulting from a poor choice of statistical technique.

Since prospectively determined reimbursement guidelines (or limits) function as surrogate prices in the market for hospital services, any conclusions regarding the appropriate choice of criteria draw heavily on economic price theory. It was concluded that three types of variables can be identified: those that are outside the control of the hospital (e.g. input prices), those that are within the control of the hospital (e.g. length of stay), and those variables that are of specific policy interest (e.g. the extent of teaching programs, the quality of care). In order to assure the long run financial health of efficient institutions in the industry, variables of the first type must be included in the set of classification criteria. Inclusion of variables of the second type (either directly or as surrogates for the type 1 variables) can be expected to promote changes in institutional behavior contrary to the cost containment objectives of the system.

With respect to the policy variables, the inclusion or exclusion decision must be made on the basis of the desires of the policy makers and the other policy instruments at hand. Thus, a program intended to encourage teaching programs should include a teaching variable in the similarity criteria so that an

institution would never be penalized for the additional cost of efficiently-run teaching programs. A program uninterested in the promotion of above average quality of care would exclude an explicit quality variable, and rely on licensure standards or other constraints to monitor the average.

Further, the specific set of classification variables that is appropriate depends on not only program policy but also on other elements of program design. The question of which hospitals should be grouped together cannot be answered without first answering the questions: What is the proposed payment or review unit (e.g., per service, per day, per admission, per unit of time)? What costs are the targets of the control program (e.g., routine costs, total costs)? How is the rate or limit to be determined within groups?

With respect to variable weighting, it was concluded that the appropriate theoretical answer is of little practical use: variables that affect cost should be weighted in the classification system according to their coefficients in the "efficient" industry cost function. Given that the parameters of this function are not known (and can be estimated only with data on existing industry practice), this result only provides guidance for a very imperfect second best approach.

Finally, the issue of system validation was discussed. Because the methodology proposed in this study attempts to classify hospitals based on efficient performance rather than actual performance, measures of statistical tidiness of the resulting hospital groups will be good validation tools of this system only if the industry is currently operating close to the optimal position.

II. Statistical Methodology

Chapter 3 began with a discussion of the selection of the grouping methodology. Cluster analysis was selected as the appropriate grouping methodology on two counts. First, unlike AID, discriminant analysis or regression analysis, cluster analysis does not require the specification of a dependent variable. Further, unlike regression analysis, factor analysis or discriminant analysis, this technique does not impose a particular functional form on the relationship among the variables.

A major conclusion of this section of the study was that the preparation of the input variables for the cluster analysis must proceed with care. While the discussion in Chapter 2 could offer only weak guidelines for the assignment of appropriate weights to the input variables, it was observed in Chapter 3 that implicit (and unknown) weights may be assigned unintentionally if the variables are multicollinear. The suggested method of removing these implicit weights was to factor analyze the input variables and use only the standardized factor scores in the computation of the similarity matrix. In this way, explicit (and known) weights can be added, and the sensitivity of the resulting group structure to changes in these weights can be tested.

After concluding that agglomerative hierarchical clustering algorithms avoid the computational problems of divisive algorithms and provide more information

than iterative algorithms, several well-known agglomerative algorithms were discussed. Since it is known that various algorithms tend to search for clusters of different sizes (e.g., the complete linkage algorithm tends to find clusters of spherical shap while the single linkage algorithm tends to find serpentine clusters), a conservative method of combining the results from a number of individual algorithms was developed. This composite approach also offered the advantage that no single algorithm had to be selected and justified--a tenuous task in most cases as empirically demonstrated in the fourth chapter.

A method was also developed for analyzing the composite dendrogram resulting from this combined approach. It was shown how a natural measure called expected distinctiveness could be used for comparing partitions from the composite dendrogram. An easily implemented algorithm based on expected distinctiveness was explained and illustrated. This algorithm can be used in choosing among the various partitions found.

Finally, in Chapter Three, a statistic for measuring the similarity between group structures was developed. This statistic, based on the null hypothesis of random grouping and equally likely groups, was useful for statistically testing the sensitivity of the classification system to changes in various parameters and procedures.

III. Empirical Results

It was stated at the beginning of this chapter that the data set available in this study precluded a definitive test of the characteristics of the classification system proposed here. However, a number of empirical results were obtained that have interesting implications for further conceptual development, data collection, and empirical analysis.

A brief summary of the empirical approach is presented here, followed by a discussion of the most useful empirical results.

The first step in the empirical analysis was to identify measures within the available data set that most closely captured the spirit of the grouping variables selected in Chapter 2. The most severe problem was with the case mix variable. The solution finally adopted was to choose two alternative approaches (the "endogenous" approach and the "exogenous" approach), both of which violated the principles established in Chapter 2. Unfortunately, in the absense of more direct measures no estimate of the bias introduced by these approaches was possible.

The input measures from these two approaches plus a third (representing the essense of the current HCFA approach) were factor analyzed. The factor scores were then assigned weights. Once again, the absense of an optimal approach resulted in the choice of two alternatives: the assignment of unit weights to each variable,[49] and the assignment of regression weights. In the latter case, the weights represented the coefficients of each factor in a regression with cost per case as the dependent variable. As noted above, this provides an empirical estimate of the cost function.

---

[49] For variables represented by more than one factor, each factor was assigned an equal proportion of the unit weight.

Cluster analysis was then performed on six sets of measures representing the combination of three variable sets (endogenous, exogenous, and HCFA) and two weighting schemes. The analysis was run first on a sample of 194 hospitals, combining the results of six algorithms into a composite dendrogram. In the analysis of the complete sample of 1070 hospitals, only three variable sets were analyzed: endogenous with regression weights, endogenous with unit weights, and exogenous with regression weights. Three algorithms were combined in the analysis of the first variable set; only the average linkage between groups algorithm was used with the latter two. Finally, a number of tests were performed on the resulting group structures to investigate their statistical characteristics.

A number of notable conclusions were suggested by these results. First, the regression analysis that was performed to estimate factor score weights was only moderately successful. The explanatory power of the equation was never as high as 50 percent, and some of the signs of the coefficients were counter-intuitive. It is possible that this was primarily a data problem, but further analysis is warranted here.

Second, the absolute cophenetic correlation coefficients calculated for the six individual algorithms indicate that no single algorithm consistently outperforms any other. Thus, the use of a composite dendrogram appears to be supported.

In general, the measure of expected distinctiveness proved to be a good indicator of partition "quality." In addition, it appeared to provide a fairly accurate measure of the ability of the partitions to be discriminated by linear functions: the correlation between the expected distinctiveness measure and the percentage of hospitals correctly classified by a linear discriminant analysis was consistently high for both the small sample and the full sample results. In both cases, the hospital clusters appear to be well defined by linear functions. The latter results suggest that further analysis is warranted to investigate the possibility that the entire hospital population might successfully be classified based on the discriminant functions determined from a sample. It is imperative to note, however, that these results may be specific to this data set and may not hold when improved variable measures and more current data are used.

Analysis of variance tests were performed to determine the sensitivity of the final group structure to changes in the factor score weights. In the small sample, the ANOVA results verified that factors assigned a zero weight in the regression analysis (i.e., those factors whose coefficients were statistically insignificant) had no effect on final group structure while factors with positive weights all had a significant impact on the outcome. This result suggests that the clusters are somewhat sensitive to weight changes, although a more systematic analysis would have to be performed to determine the extent to which this is true.

The small sample ANOVA results were supported by the full sample analysis where a strong positive relationship was found between the weights attached to each factor and their respective F values.

A partial test of cluster sensitivity to changes in the variables (as opposed to variable weights) is afforded by the partition comparison measure, Λ. When the partitions produced by the endogenous approach and the exogenous approach (both with regression weights) were compared using the Λ statistic, a reasonably high degree of similarity was found, although the degree of similarity was affected somewhat by the number of groups being compared.

The partition comparison measure was also used to provide a second test of sensitivity to changes in variable weights. The results of this test were similar to those of the previous partition comparison tests, and thus inconsistent with the ANOVA results. Further investigation is necessary to resolve this apparent inconsistency.

Finally, the final group structures suggest the presence of a number of hospitals that repeatedly fall into groups by themselves. The number of these isolates is small (approximately two percent of the complete sample), and a detailed analysis of their characteristics suggested that they indeed represent special cases (an example is a large teaching institution that is located in a rural area). To the extent that these hospitals are atypical, it is desirable that the classification system respond accordingly.

IV. Directions for Further Research

The purpose of this study was to develop the conceptual and methodological underpinnings for a hospital classification system useful in a prospective rate determination or review program. In the process of this development, and in the initial tests of its feasibility, a number of questions have arisen that warrant further investigation.

The most obvious extension is the development of better empirical measures of the variables. Case mix and case mix severity are at the top of the list, but the measures for input prices and rural markets used in this study also leave room for improvement. Once better measures are obtained for a sample of hospitals, new clusters can be generated and compared with the present results. These comparisons will then be helpful in making decisions about the benefits of larger-scale data collection efforts for this purpose.

Further investigation of the sensitivity of the clusters to changes in variable weights is also appropriate. The tests reported here are only suggestive, since their results appear to be somewhat inconsistent.

Cluster stability over time is important for the long term maintenance of a payment or review system based on hospital classification methods. Since the data set available to the study contained only 1973 figures, questions of cluster stability over time could not be addressed.

The importance of a reliable measure of partition comparison in these extensions is obvious. While such a measure was developed and used in this study, future work should investigate other possibilities and compare them with the study statistic.

V. Concluding Remarks

It must be remembered that the original rationale for developing this hospital classification system was as an integral part of a prospective payment determination or review program. By design, this study has not addressed the important issue of exactly how this integration takes place. That is, how would hospital clusters be used to set or review rates? Upon what payment unit should payment or review be based? Which costs should be included? All of these questions must be answered before an effective program can be implemented.

The output characteristics of an industry, its long run stability, and its response to technological change are all determined by how the industry is financed. The institution of a nationwide prospective payment determination or review system for even a segment of the hospital industry represents a substantial change from current practice. If this change is to be positive and to achieve the objectives implicit in its development, it must be carefully planned and carefully implemented. It is hoped that this study can provide a first step in that direction.

APPENDIX A

TABLE A.1.1.

DESCRIPTIVE MEASURES BY GROUP: ENDOGENOUS APPROACH - REGRESSION WEIGHTS

Measure	Group 1 Mean	Group 2 Mean	Group 3 Mean	Group 4 Mean	Group 5 Mean	Group 6 Mean	Group 7 Mean
Manufacturing Wage	4.44	2.98	3.18	2.59	2.29	4.02	5.04
Public Wage	4.86	3.60	3.82	2.69	3.37	4.23	5.82
Retail Wage	2.43	1.88	1.98	1.95	1.72	2.13	3.23
Hospital Wage	7,543.07	6,030.39	6,014.28	5,538.71	19,564.00	7,058.68	10,396.24
UPOI	.928	.733	.807	.859	.630	.932	.985
Basic Services	3.91	3.81	3.85	4.00	4.00	3.95	4.00
Quality Services	5.06	2.64	3.64	5.00	3.00	5.85	5.00
Complex Services	8.39	1.49	2.98	5.00	.00	15.30	9.00
Community Services	4.37	.76	.66	1.00	.00	12.95	11.00
Birth/Patient	.09	.09	.08	.16	.05	.10	.19
Surgeries /Patient	.51	.21	.29	.63	.24	.47	.46
Outvisit /Patient	4.38	3.41	2.89	2.43	.81	9.47	6.96
Severity Ages	.203	.232	.228	.216	.243	.187	.206
Income LT $4000	.140	.282	.260	.400	.528	.168	.152
Number of Cases	737	174	53	1	1	20	3
Number of Beds	312.38	51.69	81.28	234.00	25.00	560.40	339.33
Percentage Accredited	.948	.431	.811	1.000	.000	1.000	1.000
Percentage Med School	.305	.023	.000	.000	.000	.950	.667
Percentage in SMSA Area	.737	.040	.038	1.000	.000	.750	1.000

TABLE A. 1.1.
(cont.)

Measure	Group 8 Mean	Group 9 Mean	Group 10 Mean	Group 11 Mean	Group 12 Mean	Group 13 Mean
Manufacturing Wage	4.55	4.14	4.83	4.82	5.41	3.76
Public Wage	4.53	4.76	5.49	5.42	5.43	4.34
Retail Wage	2.50	2.40	3.01	2.74	2.89	2.29
Hospital Wage	11,921.00	8,048.62	9,525.62	9,307.91	9,805.25	9,989.86
UPOI	.925	.791	.975	.964	.880	.922
Basic Services	4.00	4.00	4.00	4.00	3.50	4.00
Quality Services	1.00	5.00	5.33	5.80	4.25	2.00
Complex Services	.00	5.00	12.67	14.12	4.50	3.00
Community Services	.00	11.33	14.00	14.17	1.75	6.00
Birth/Patient	.13	.09	.08	.11	.15	.00
Surgeries/Patient	.11	.38	.37	.52	1.37	.00
Outvisit/Patient	15.88	11.25	39.54	11.69	1.65	146.91
Severity Ages	.178	.206	.196	.201	.201	.167
Income LT $4000	.151	.121	.116	.124	.116	.136
Number of Cases	1	3	3	69	4	1
Number of Beds	18.00	142.00	830.00	621.20	84.75	28.00
Percentage Accredited	.000	1.000	1.000	.986	1.000	.000
Percentage Med School	.000	.333	.333	.884	.000	.000
Percentage in SMSA Area	.000	.000	1.000	.971	1.000	1.000

TABLE A.1.2.

Descriptive Measures by Group: Endogenous Approach - Regression Weights -- 1st Suboptimal Group Structures

Measure	Group 101 Mean	Group 102 Mean	Group 103 Mean	Group 104 Mean	Group 105 Mean	Group 106 Mean	Group 107 Mean
Manufacturing Wage	4.38	4.04	5.09	4.90	4.70	3.98	3.80
Public Wage	4.72	4.45	5.97	5.69	5.48	4.30	3.91
Retail Wage	2.37	2.33	3.26	2.28	2.52	2.27	2.14
Hospital Wage	7274.62	6801.92	8019.09	8045.37	8081.13	6435.75	6356.57
UPOI	.920	.937	.968	.926	.961	.934	.935
Basic Services	3.92	4.00	1.00	4.00	3.90	3.96	4.00
Quality Services	5.08	5.65	4.00	5.00	5.50	5.73	5.75
Complex Services	7.58	13.48	12.00	10.00	12.70	11.71	13.75
Community Services	3.34	8.70	8.00	4.00	6.50	3.60	9.25
Birth/Patient	.10	.09	.00	.11	.09	.08	.10
Surgeries/Patient	.51	.50	.50	.54	.62	.55	.51
Outvisit/Patient	4.10	3.87	4.69	6.58	3.48	2.06	4.49
Severity Ages	.197	.205	.211	.203	.207	.219	.180
Income LT $4000	.138	.151	.142	.096	.118	.188	.196
Number of beds	166.74	290.93	.00	.00	119.64	186.40	260.34
Percentage Accredited	.221	.000	.000	.000	.000	.000	.000
Percentage Medical School	.422	.507	.000	.000	.466	.483	.577
Percentage in SMSA area	.462	.422	.000	.000	.000	.425	.000

TABLE A.1.2.

1st Suboptimal Group Structures (Continued)

Measure	Group 108 Mean	Group 109 Mean	Group 110 Mean	Group 111 Mean	Group 112 Mean	Group 113 Mean	Group 114 Mean	Group 115 Mean
Manu. Wage	3.25	4.88	4.97	3.48	3.65	4.01	4.87	4.44
Public Wage	3.79	5.77	5.52	4.06	3.93	4.45	5.25	4.95
Retail Wage	2.08	2.85	2.70	2.17	2.22	2.33	2.54	2.31
Hosp. Wage	5819.72	9758.70	8488.71	6809.94	6414.63	7283.83	8307.39	7122.17
UPOI	.871	.971	.962	.845	.864	.870	.952	.833
Basic Services	3.88	4.00	3.77	3.88	3.93	3.75	3.99	3.88
Quality Serv.	5.71	5.54	4.60	2.88	5.06	1.67	5.66	4.50
Complex Serv.	9.76	8.85	5.38	1.50	7.00	.83	12.18	3.50
Commun. Serv.	1.94	8.15	2.37	1.28	1.93	1.67	8.91	1.25
Birth/Patient	.09	.11	.09	.08	.10	.08	.10	.08
Surg./Patient	.52	.48	.52	.23	.50	.16	.54	.68
Outvisit/Patient	2.23	7.65	4.45	5.89	3.15	7.33	5.75	3.17
Severity Ages	.223	.203	.201	.220	.213	.206	.201	.208
Income LT $4000	.287	.104	.115	.210	.176	.139	.111	.126
Number beds	114.57	136.20	113.26	39.43	69.28	34.69	197.33	57.54
Percentage Acc.	.000	.000	.219	.457	.363	.492	.091	.354
Percentage Med School	.000	.480	.338	.000	.000	.000	.474	.000
Percentage in SMSA area	.437	.000	.256	.177	.363	.492	.234	.535

TABLE A.1.3.

Descriptive Measures by Group: Endogenous Approach - Regression Weights -- 2nd Suboptimal Group Structures

Measure	Group 10101 Mean	Group 10102 Mean	Group 10103 Mean	Group 10104 Mean	Group 10105 Mean
Manufacturing Wage	5.37	4.39	4.61	4.35	4.14
Public Wage	5.65	4.83	4.99	4.40	4.61
Retail Wage	2.73	2.36	2.39	2.29	2.35
Hospital Wage	9047.77	7275.89	7304.43	7132.31	6565.79
UPOI	.973	.940	.908	.900	.939
Basic Services	4.00	3.98	4.00	4.00	3.84
Quality Services	5.50	5.64	5.71	5.33	5.32
Complex Services	9.50	10.66	10.29	9.08	8.04
Community Services	6.50	4.60	5.57	5.42	1.64
Birth/Patient	.13	.10	.09	.09	.08
Surgeries/Patient	.55	.56	.48	.48	.64
Outvisit/Patient	4.35	3.81	5.96	6.35	1.84
Severity Ages	.196	..202	.206	.188	.199
Income LT $4000	.104	.136	.142	.133	.148
Number of beds	401.57	431.76	329.43	321.17	310.32
Percentage Accredited	1.000	1.000	1.000	1.000	1.000
Percentage Med School	.429	.400	.286	.333	.280
Percentage in SMSA area	1.000	.860	.571	.583	.760

TABLE A.1.3.

2nd Suboptimal Group Structures (Continued)

Measure	Group 10106 Mean	Group 10107 Mean	Group 10108 Mean	Group 10109 Mean
Manufacturing Wage	3.73	4.84	4.46	4.05
Public Wage	3.99	5.20	4.65	4.40
Retail Wage	2.16	2.50	2.42	2.27
Hospital Wage	6439.55	7765.14	7344.12	7049.51
UPOI	.896	.955	.935	.870
Basic Services	4.00	3.85	3.89	3.91
Quality Services	5.53	5.16	3.79	4.56
Complex Services	9.50	7.92	4.05	4.54
Community Services	3.83	3.53	2.32	1.91
Birth/Patient	.10	.10	.07	.10
Surgeries/Patient	.49	.57	.41	.42
Outvisit/Patient	4.25	4.11	5.51	4.05
Severity Ages	.184	.193	.195	.204
Income LT $4000	.181	.107	.123	.157
Number of beds	295.13	295.38	131.00	138.16
Percentage Accredited	1.000	.946	.842	.889
Percentage Med School	.167	.338	.000	.037
Percentage in SMSA area	.567	.932	.789	.346

TABLE A.2.1.

Descriptive Measures by Group: Endogenous Approach - Unit Weights
Optimal Group Structures - (N = 1070 Hospitals)

Measure	Group 1 Mean	Group 2 Mean	Group 3 Mean	Group 4 Mean	Group 5 Mean	Group 6 Mean
Manufacturing Wage	4.47	4.59	3.29	3.58	5.23	5.88
Public Wage	4.85	5.52	3.77	4.30	5.56	5.85
Retail Wage	2.42	2.97	2.15	2.31	.25	3.05
Hospital Wage	7541.06	10480.23	7942.89	6958.98	6490.34	9438.66
UPOI	.928	.976	.875	.899	.942	.974
Basic Services	3.92	4.00	4.00	3.83	4.00	3.00
Quality Services	5.09	5.52	2.00	5.33	6.00	3.50
Complex Services	8.74	12.36	3.00	7.17	15.00	2.50
Community Services	5.00	12.76	2.00	4.17	11.00	2.00
Births/Patient	.10	.10	.77	.09	.10	.23
Surgeries/Patient	.50	.55	.24	.49	.50	1.44
Outvisit/Patient	4.82	12.37	6.75	6.02	4.68	.79
Severity Ages	.199	.211	.186	.328	.182	.206
Income LT $4000	.135	.157	.211	.228	.182	.127
Number of beds	326.77	590.10	42.00	333.42	843.00	60.00
Percentage Accredited	.943	1.00	1.00	.833	1.00	1.00
Percentage Med School	.349	.714	.000	.083	1.00	.000
Percentage in SMSA area	.740	1.000	.000	.500	1.000	1.000

TABLE A.2.1.

Optimal Group Structures (Continued)

Measure	Group 7 Mean	Group 8 Mean	Group 9 Mean	Group 10 Mean	Group 11 Mean	Group 12 Mean
Manufacturing Wage	5.09	1.98	2.99	4.09	3.76	2.29
Public Wage	5.97	3.81	3.68	4.89	4.34	3.37
Retail Wage	3.26	1.65	1.92	2.69	2.29	1.72
Hospital Wage	8019.09	5436.76	6079.88	8936.29	9989.86	19564.00
UPOI	.968	.621	.752	.612	.922	.630
Basic Services	1.00	3.00	3.84	4.00	4.00	4.00
Quality Services	4.00	1.25	2.89	6.00	2.00	3.00
Complex Services	12.00	.00	2.06	8.00	3.00	.00
Community Services	8.00	.25	.86	.00	6.00	.00
Births/Patient	.00	.08	.09	.09	.00	.05
Surgeries/Patient	.50	.08	.24	.98	.00	.24
Outvisit/Patient	4.69	7.99	3.39	2.34	146.91	.81
Severity Ages	.211	.252	.234	.218	.167	.243
Income LT $4000	.142	.520	.285	.181	.136	.528
Number of beds	333.00	34.50	65.89	132.00	28.00	25.00
Percentage Accredited	1.000	.000	.546	1.000	.000	.000
Percentage Med School	1.000	.000	.023	.000	.000	.000
Percentage in SMSA area	1.000	.000	.042	1.000	1.000	.000

TABLE A.2.2.

Descriptive Measures by Group: Endogenous Approach - Unit Weights
1st Suboptimal Group Structures (N = 1070 Hospitals)

Measure	Group 901 Mean	Group 902 Mean	Group 903 Mean
Manufacturing Wage	2.98	5.17	2.12
Public Wage	3.67	5.26	3.76
Retail Wage	1.92	2.68	1.59
Hospital Wage	6066.65	7025.32	7965.58
UPOI	.751	.753	.828
Basic Services	3.84	4.00	4.00
Quality Services	2.86	6.00	5.00
Complex Services	1.99	16.00	3.00
Community Services	.77	15.00	6.00
Births/Patient	.08	.08	.45
Surgeries/Patient	.24	.65	.21
Outvisit/Patient	3.32	15.41	4.95
Severity Ages	.234	.208	.216
Income LT $4000	.285	.138	.417
Number of Beds	63.47	530.00	120.00
Percentage Accredited	.542	1.00	1.00
Percentage Med School	.014	1.000	1.000
Percentage in SMSA Area	.037	1.000	.000

TABLE A.2.3.

Descriptive Measures by Group: Endogenous Approach - Unit Weights
2nd Suboptimal Group Structures (N = 1070 Hospitals)

Measures	Group 90101 Mean	Group 90102 Mean	Group 90103 Mean
Manufacturing Wage	2.96	2.65	3.58
Public Wage	3.69	3.55	3.84
Retail Wage	1.99	1.76	1.99
Hospital Wage	6040.94	6024.36	6208.39
UPOI	.810	.679	.698
Basic Services	3.90	3.77	3.74
Quality Services	3.34	2.08	2.72
Complex Services	2.81	.69	1.69
Community Services	.93	.45	.82
Births/Patient	.09	.07	.09
Surgeries/Patient	.28	.18	.22
Outvisit/Patient	3.51	3.02	3.26
Severity Ages	.236	.245	.212
Income LT $4000	.294	.323	.197
Number of Beds	85.74	38.58	38.54
Percentage Accredited	.708	.339	.385
Percentage Med School	.009	.016	.026
Percentage in SMSA Area	.044	.000	.077

TABLE A.2.4.

Descriptive Measures by Group: Endogenous Approach - Unit Weights
3rd Suboptimal Group Structures (N = 1070 Hospitals)

Measure	Group 101 Mean	Group 102 Mean	Group 103 Mean	Group 104 Mean	Group 105 Mean	Group 106 Mean
Manufacturing Wage	4.95	4.62	4.37	2.91	4.63	4.04
Public Wage	5.32	5.45	4.97	3.50	3.89	4.42
Retail Wage	2.58	2.66	2.45	1.93	2.18	2.27
Hospital Wage	8193.22	7142.66	7258.09	4704.57	5342.26	6806.23
UPOI	.955	.978	.918	.861	.883	.930
Basic Services	3.94	3.00	3.15	3.00	4.00	3.98
Quality Services	5.28	4.83	1.38	2.71	5.50	5.64
Complex Services	9.34	4.83	1.54	2.71	12.00	11.28
Community Services	5.75	4.67	.77	.57	2.00	5.90
Births/Patient	.10	.42	.04	.06	.05	.10
Surgeries/Patient	.55	.51	.25	.23	.42	.52
Outvisit/Patient	5.16	4.01	3.74	1.83	1.38	4.59
Severity Ages	.196	.200	.197	.203	.180	.197
Income LT $4000	.106	.129	.138	.241	.121	.163
Number of Beds	359.25	114.50	41.62	78.57	235.00	432.92
Percentage Accredited	.975	.833	.692	.571	1.000	.991
Percentage in SMSA Area	.947	1.00	.692	.143	.500	.764
Percentage in Med School	.425	.333	.000	.000	.000	.444

TABLE A.2.4

3rd Suboptimal Group Structures (Continued)

Measure	Group 107 Mean	Group 108 Mean	Group 109 Mean	Group 110 Mean	Group 111 Mean
Manufacturing Wage	3.73	3.72	4.00	3.86	4.39
Public Wage	4.25	4.52	3.83	3.99	4.84
Retail Wage	2.20	2.32	2.17	2.09	2.43
Hospital WAge	6738.86	6906.00	6946.47	6872.81	8625.72
UPOI	.890	.877	.859	.790	.807
Basic Services	3.69	3.97	3.92	4.00	4.00
Quality Services	2.56	5.24	5.44	4.20	4.00
Complex Services	3.17	5.67	5.96	3.17	4.20
Community Services	1.50	3.27	3.20	.83	3.47
Births/Patient	.06	.10	.12	.10	.10
Surgeries/Patient	.29	.45	.45	.33	.44
Outvisit/Patient	5.06	4.09	5.06	3.78	5.23
Severity Ages	.210	.223	.188	.221	.208
Income LT $4000	.163	.190	.154	.173	.127
Number of Beds	87.78	190.55	155.36	83.29	119.40
Percentage Accredited	.667	.939	1.000	.714	.933
Percentage Med School	.056	.030	.040	.029	.133
Percentage in SMSA Area	.278	.091	.120	.057	.267

TABLE A.3.1

Descriptive Measures by Group: Exogenous Approach - Regression Weights
Optimal Group Structures

Measure	Group 1 Mean	Group 2 Mean	Group 3 Mean]	Group 4 Mean	Group 5 Mean
Manufacturing Wage	4.45	2.93	4.81	4.63	4.09
Public Wage	4.87	3.65	3.90	5.05	4.89
Retail Wage	2.42	1.92	1.67	2.57	2.69
Hospital Wage	7662	5990	5222	8018	8983
UPOI	.925	.756	.618	.954	.612
Family Income	9885	6459	5355	9426	10738
Labor Rate	.574	.517	.224	.583	.526
Disable Rate	.095	.124	.076	.095	.099
% M.D.s 60+	.20	.25	.33	.13	.29
ObGyns/10,000	.087	.020	.000	.195	.088
Primary/10,000	.772	.481	.264	1.845	1.125
Medical Specialist/10,000	2.968	.473	.000	10.741	4.170
Other Direct Care Specialists	.248	.041	.000	.908	.960
Other Specialists/10,000	.199	.030	.000	.753	.872
Surgeons/10,000	.401	.107	.100	1.084	.581
High Demand Ages	.411	.416	.242	.428	.411
% Families LT $4000	.134	.295	.380	.152	.131
Number of Beds	322.32	71.52	23.00	436.93	132.00
Percentage Accredited	.936	.548	.000	.977	1.000
Percentage Med School	.331	.018	.000	.773	.000
Percentage in SMSA Area	.737	.046	.000	.909	1.000

TABLE A. 3.1.

Descriptive Measures by Group: Exogenous Approach - Regression Weights

Optimal Group Structures (N = 1070 Hospitals) (Continued)

Measure	Group 6 Mean	Group 7 Mean	Group 8 Mean	Group 9 Mean	Group 10 Mean
Manufacturing Wage	5.11	3.83	4.26	2.86	6.65
Public Wage	5.72	4.06	6.79	3.53	4.04
Retail Wage	3.37	2.31	2.01	2.10	2.10
Hospital Wage	10240	7156	7028	7688	8156
UPOI	.982	.927	.894	.898	.944
Family Income	8983	9089	8267	8770	10972
Labor Rate	.597	.587	.486	.561	.653
Disable Rate	.102	.078	.052	.059	.075
Percentage M.D.s 60+	.24	.03	.08	.05	.06
ObGyns/10,000	.397	.278	.424	.347	.392
Primary Care/10,000	3.363	2.849	3.029	3.500	6.373
Medical Specialists/10,000	19.835	19.152	19.990	23.394	54.337
Direct Care Specialists/10,000	2.011	1.399	1.757	2.634	3.365
Other Specialists/10,000	1.792	1.239	1.636	1.785	2.556
Surgeons/10,000	1.933	2.122	3.332	1.872	4.114
High Demand Ages	..443	.419	.432	.412	.439
% Families LT $4000	.177	.152	.161	.163	.102
Number of Beds	699.71	621.00	428.00	467.00	789.50
Percentage Accredited	1.000	1.000	1.000	1.000	1.000
Percentage Med School	.643	.750	1.000	1.000	1.000
Percentage in SMSA Area	1.000	.250	.000	1.000	1.000

TABLE A.3.2

Descriptive Measures by Group: Exogenous Approach - Regression Weights
1st Suboptimal Group Structures (N = 1070 Hospitals)

Measures	Group 401 Mean	Group 402 Mean	Group 103 Mean	Group 404 Mean
Manufacturing Wage	4.08	5.14	5.29	4.56
Public Wage	4.30	5.47	5.81	5.87
Retail Wage	2.32	2.87	2.59	2.59
Hospital Wage	6857	8553	10817	8026
UPOI	.936	.969	.963	.973
Family Income	8850	9852	10486	9280
Labor Rate	.563	.606	.590	.582
Disable Rate	.090	.103	.083	.099
% M.D.s 60+	.10	.16	.11	.14
ObGyns/10,000	.160	.237	.211	.189
Primary Care Specialists/10,000	1.378	1.980	2.394	2.658
Medical Specialists/10,000	7.073	11.548	15.343	17.655
Direct Care Specialists/10,000	.683	.936	1.295	1.296
Other Specialists/10,000	.564	.776	1.069	1.087
Surgeons/10,000	.931	1.127	1.299	1.323
High Demand Ages	.417	.430	.440	.448
% Families LT $4000	.166	.143	.125	.154
Number of beds	489.05	370.53	560.00	315.00
Percentage Accredited	.947	1.000	1.000	1.000
Percentage Med School	.842	.600	1.000	.800
Percentage in SMSA Area	.789	1.000	1.000	1.000

TABLE A.3.3.

Descriptive Measures by Group: Exogenous Approach - Regression Weights
2nd Suboptimal Group Structures (N = 1070 Hospitals)

Measures	Group 201 Mean	Group 202 Mean	Group 203 Mean
Manufacturing Wage	2.41	3.62	2.88
Public Wage	3.57	4.07	3.60
Retail Wage	1.71	2.05	1.92
Hospital Wage	6579	6687	5856
UPOI	.648	.680	.774
Family Income	4518	9151	6374
Labor Rate	.493	.551	.514
Disable Rate	.144	.113	.124
% M.D.s 60+	.30	.34	.24
ObGyns/10,000	.000	.017	.022
Primary Care Specialists/10,000	.398	.430	.494
Medical Specialists/10,000	.171	.152	.537
Direct Care Sepcialists/10,000	.013	.024	.045
Other Specialists/10,000	.010	.020	.032
Surgeons/10,000	.054	.060	.117
High Demand Ages	.415	.412	.416
% Families LT $4000	.464	.187	.297
Number of Beds	33.77	40.83	78.26
Percentage Accredited	.231	.417	.588
Percentage Med School	.000	.042	.016
Percentage in SMSA Area	.000	.125	.038

TABLE A.3.4

Descriptive Measures by Group: Exogenous Approach - Regression Weights

3rd Suboptimal Group Structures (N = 1070 Hospitals)

Measures	Group 101 Mean	Group 102 Mean	Group 103 Mean
Manufacturing Wage	4.87	3.80	5.17
Public Wage	5.23	4.30	5.26
Retail Wage	2.55	2.22	2.68
Hospital Wage	8189	6835	7025
UPOI	.95	.885	.753
Family Income	10840	8387	9210
Labor Rate	.594	.543	.604
Disable Rate	.090	.102	.106
% M.D.s 60+	.18	.23	.12
ObGyns/10,000	.102	.063	.118
Primary Care/10,000	.837	.670	.995
Medical Specialists/10,000	3.571	2.013	5.245
Direct Care Specialists/10,000	.298	.168	.391
Other Specialists/10,000	.244	.129	.284
Surgeons/10,000	.444	.333	.535
High Demand Ages	.409	.415	.418
% Families LT $4000	.107	.177	.138
Number of Beds	368.68	248.75	530.00
Percentage Accredited	.957	.888	1.000
Percentage Med School	.416	.194	1.000
Percentage in SMSA Area	.933	.428	1.000

APPENDIX B

APPENDIX B

CORRELATION MATRIX FOR ENDOGENOUS MEASURES

	MFGW	PUBW	RETW	HOSPW	POOR	BASIC	QUALTY	COMPLX	COMMUN	BIRTHD	SURGPD	OUTPPD	HIDEMN
MFGW	1.00	0.62	0.56	0.44	-0.69	0.10	0.43	0.48	0.41	0.07	0.48	0.12	-0.39
PUBW		1.00	0.70	0.48	-0.62	0.03	0.43	0.50	0.45	0.08	0.47	0.15	-0.34
RETW			1.00	0.53	-0.57	0.06	0.45	0.49	0.46	0.09	0.49	0.17	-0.36
HOSPW				1.00	-0.48	0.11	0.27	0.35	0.43	0.10	0.38	0.21	-0.19
POOR					1.00	-0.12	-0.46	-0.50	-0.43	-0.07	-0.53	-0.13	0.49
BASIC						1.00	0.35	0.25	0.20	0.05	0.16	0.10	-0.08
QUALTY							1.00	0.75	0.55	0.17	0.56	0.10	-0.33
COMPLX								1.00	0.77	0.10	0.57	0.17	-0.36
COMMUN									1.00	0.14	0.41	0.31	-0.26
BIRTHD										1.00	0.10	0.04	-0.10
SURGPD											1.00	0.01	-0.33
OUTPPD												1.00	-0.08
HIDEMN													1.00

MFGW - Manufacturing Hourly Wage
PUBW - Transportation & Public Utilities Hourly Wage
RETW - Retail Hourly Wage
HOSPW - Hospital Yearly Wage
POOR - Families Earning Less Than $4,000 Yearly
BASIC - Number of Basic Services
QUALTY - Number of Quality Enhancing Services
COMPLX - Number of Complex Services
COMMUN - Number of Community Services
BIRTHD - Ratio of Births to Total Discharges
SURGPD - Ratio of Surgical Operations to Total Discharges
OUTPPD - Ratio of Outpatient Visits to Total Discharges
HIDEMN - High Demand Age

APPENDIX B

CORRELATION MATRIX FOR EXOGENOUS MEASURES

	MFGW	PUBW	RETW	HSPW	DEMAN	POOR	FAMIN	LABOR	OBGYN	PRIMC	NONWT	DISAB	MD60+	MEDSP	DIRCR	OTHSP	SURSP
MFGW	1.00	0.64	0.56	0.44	-0.15	-0.69	0.74	0.45	0.44	0.30	0.24	-0.40	-0.21	0.35	0.36	0.35	0.36
PUBW		1.00	0.70	0.48	-0.05	-0.62	0.69	0.50	0.53	0.38	0.09	-0.33	-0.20	0.40	0.40	0.41	0.42
RETW			1.00	0.53	-0.02	-0.57	0.66	0.46	0.65	0.48	0.22	-0.25	-0.24	0.49	0.52	0.53	0.50
HSPW				1.00	-0.06	-0.48	0.56	0.33	0.38	0.35	-0.01	-0.33	-0.09	0.35	0.38	0.38	0.32
DEMAN					1.00	0.22	-0.25	-0.13	0.15	0.24	0.06	0.23	0.10	0.22	0.21	0.21	0.21
POOR						1.00	-0.92	-0.68	-0.36	-0.22	0.29	0.65	0.22	-0.26	-0.28	-0.27	-0.29
FAMIN							1.00	0.68	0.40	0.25	-0.16	-0.60	-0.26	0.30	0.31	0.31	0.30
LABOR								1.00	0.37	0.26	0.02	-0.44	-0.26	0.28	0.28	0.27	0.29
OBGYN									1.00	0.85	0.33	-0.24	-0.31	0.84	0.86	0.86	0.91
PRIMC										1.00	0.20	-0.15	-0.19	0.97	0.95	0.94	0.94
NONWT											1.00	0.24	-0.14	0.21	0.18	0.18	0.23
DISAB												1.00	0.16	-0.19	-0.21	-0.20	-0.21
MD60+													1.00	-0.24	-0.23	-0.22	-0.28
MEDSP														1.00	0.95	0.94	0.95
DIRCR															1.00	0.99	0.93
OTHSP																1.00	0.92
SURSP																	1.00

MFGW - Manufacturing Hourly Wage
PUBW - Transportation & Public Utilities Hourly Wage
RETW - Retail Hourly Wage
HSPW - Hospital Yearly Wage
DEMAN - High Demand Ages
POOR - Families Earning Less Than $4,000 Yearly
FAMIN - Family Income (Yearly)
LABOR - Labor Participation Rate
OBGYN - Obstetricians/Gynecologists per 10,000 Population
PRIMC - Primary Care M.D.'s per 10,000 Population
NONWT - Percentage of Population Non-White
DISAB - Disability Rate, Ages 16-64
MD60+ - Percentage of M.D.'s Age 60 and Over
MEDSP - Medical Specialists per 10,000 Population
DIRCR - Direct Care Specialists per 10,000 Population
OTHSP - Other Specialists per 10,000 Population
SURSP - Surgical Operations per 10,000 Population

APPENDIX B

CORRELATION MATRIX FOR SSA VARIABLES

	FAMINC	TOTBED	SMSA
FAMINC	1.00	0.42	0.66
TOTBED		1.00	0.56
SMSA			1.00

FAMINC - Family Income (Yearly)
TOTBED - Total Beds
SMSA - SMSA/non-SMSA

APPENDIX C

PROGRAM SET FOR SMALL SAMPLES

(For Sets With Less Than 350 Objects/Variables)

The small sample program set currently in use completes the clustering procedure--from distance matrix computations to composite dendrogram--in one job. Control over the clustering procedure is provided by a file of parameters that is read at the first step and by alterations to the operating system control cards. The parameter file contains information on the number of entities, the number of variables, input format of the data, and optional weights used in computing distances (either Euclidean or squared Euclidean). Data input consists of the variables and an optional file of labels identifying each entity. The control cards are altered to specify such information as the names of the data files and the clustering methods to be used in finding the composite dendrogram. The actual programs are kept in compiled form on a file and are fetched and executed as needed by the procedure. The steps in a typical run are as follows:

1) Execute program EUCDP, which reads the input parameter file, prepares the file for the other job steps, and computes the matrix of distances between all pairs of entities. The distance matrix, of which only the lower-lefthand part is written, is based on the variable values and variable weights.

2) This step is executed once for each clustering method, and contains three sub-steps:

   a) Program RUNCLUS and associated subroutines perform the clustering. The method used is determined by the version of subroutine METHOD that is loaded. The program writes a vector of the joining distances for each merge.

b) Program CLUSVEC converts the joining distance vector into a full lower lefthand distance matrix.

c) System routine SORTMRG sorts the elements of the joining distance matrix so that the elements are in the proper order.

3) Program CORSTD uses the original distance matrix and the joining distance matrix from each method to produce a composite joining distance matrix. Each pairwise composite joining distance is the combination of a pair's output distances standardized and summed across all methods, weighted by the square of the Spearman correlation coefficients (i.e., the cophenetic correlation coefficient) between the original distance matrix and method joining distance matrices. The program calculates the cophenetic correlation coefficients, standardizes the matrices, and as an option can write the correlation and standardized matrices for further analysis.

4) The composite dendrogram is produced by running the complete linkage clustering method on the composite joining distance matrix.

The job that compiles the programs (file 1 of new program tape) compiles the following main programs and subroutines and stores them on a permanent file, for use by the clustering job:

EUCDP program for step 1 of cluster run

RUNCLUS
CNTRL
LFIND
CLSTR
MTXIN
TREE

} main program and common subroutines used for clustering steps 2 and 4.

METHOD single linkage METHOD complete linkage METHOD average between METHOD average within METHOD median METHOD centroid sorting METHOD Ward's	one of these subroutines is used in each execution of steps 2a and 4 to determine the clustering method. Normally we execute step 2 six times, using the complete linkage through Ward's methods, and use the complete linkage method in step 4.
CLUSVEC	program for step 2b
CORSTD SORT	main program and subroutine for step 3.

The job that performs the clustering (file 2 of new program tape) consists of operating systems control cards, the input parameters, and sets of SORT/MERGE directives for the executions of step 2c. The job is set up to read the data and entity labels from permanent (disk) files. It clusters using six methods, omitting the single linkage method, and uses the complete linkage method for the composite cluster. The composite distance matrix is saved on a permanent file. If the job terminates normally, only the composite dendrogram and merge information are printed; if the job terminates abnormally, the output (dendrograms) from all methods executed are printed and joining distance matrices are saved on permanent files.

FILE 1 OF PROGRAM TAPE -- PROGRAM COMPILATION JOB

```
JBCRL,T40,P1.
ACCOUNT.
COPYSF,INPUT,DUM,SO,COMB.
REWIND,DUM.
FTN,OPT=2,I=DUM.
REWIND,LGO.
CATALOG,LGO,RCBF2,ID=RCBF2,RP=14.
ITEMIZE,LGO.
AUDIT.
*EOR
      PROGRAM EUCOP(INPUT = 2038, TAPE1, TAPE2, OUTPUT = 2038,
     1 RCPAR = 2038, TAPE3 = RCPAR, TAPE5 = INPUT, TAPE6 = OUTPUT)
      COMMON X(1000)
      DIMENSION D(10), WT(32), RC1(7), SUS(2), IRD1(2), IRD2(2)
      INTEGER FMT(8), FMT1(2), FMT2(2), IDCS(2, 3)
      DATA IRD1, IRD2 / 2*1H , 10H(READ UNTI, 6HL EOF) /
      DATA FMT / 8H(0F10.6), 7 * 1H  /, WT / 32 * 1.0 /
      DATA FMT1 / 10H(8F10.6/0F, 5H10.6)/,
     1     FMT2 / 10H(0(8F10.6/, 9H),0F10.6) /
      DATA SUS / 8H SQUARED, 9HUNSQUARED /, MEMP / 4LMEMP /
      DATA IDCS / 10H(ALL MERGE, 2HS), 10H(AUTO-CUTO, 3HFF), 9H= HIGHEST
     1 , 5HMERGE /
C
C     COMPUTE L.L.H. TRIANGULAR MATRIX OF EUCLIDEAN DISTANCES...
C     READ FROM FILE 5 (INPUT)
C        PARAMETERS, FORMAT(6I4,F4.2,3I4,7A4)...
C           IUR .NE. 0 TO READ IN WEIGHT VECTOR, ELSE USE UNIT WTS.
C           NH = NUMBER OF ENTITIES (IF < 1, READ UNTIL EOF)
C           NV = NUMBER OF VARIABLES OR FACTORS
C           IFMT .NE. 0 TO READ IN INPUT FORMAT FOR ROW OF SCORES, ELSE
C                    USE DEFAULT FORMAT (8F10.6), FOR A ROW
C           ISQRT .NE. 0 TO OUTPUT UNSQUARED EUCLIDEAN DISTANCES
C                 .EQ. 0 OR BLANK TO OUTPUT SQUARED EUCLIDEAN DISTANCES
C           NMETH = NUMBER OF CLUSTERING METHODS (USED BY COMBVEC AND
C                   COMPDM)... DEFAULT (0 OR BLANK) IS 6
C           SAM = FRACTION SAMPLE FOR COPHENETIC CORRELATIONS...
C                 GE 1.00 USE ALL PAIRS
C                 LE 0 (DEFAULT) SELECT SAMPLE TO USE ABOUT 600 PAIRS
C           ICS .EQ. 0 OR BLANK...USE ALL MERGES TO CONSTRUCT DENDOGRAMS
C               .GT. 0...USE JOINING LEVELS UP TO THAT OF ICS-TH MERGE
C               .LT. 0...USE AUTO-CUTOFF OPTION FOR DENDOGRAM
C           ISJL .NE. 0...WRITE FILE OF COPHENETIC CORRELATIONS AND STAN
C                DARDIZED ORIGINAL AND JOINING DISTANCES ON FILE 8
C                .EQ. 0 OR BLANK...DON'T WRITE FILE 8
C           LBLFLG .EQ. 0 OR BLANK...NO ENTITY LABELS ARE SUPPLIED
C                  .NE. 0...FILE 9 IS SUPPLIED, WITH LABELS IN FORMAT
C                  (//3X,5A5,A4)
C           MNAM(1..NMETH) = OPTIONAL ARRAY OF METHOD NAMES, 1-4 CHAR-
C              ACTERS EACH, LEFT-JUSTIFIED WITH NO EMBEDDED BLANKS
C        CARD(S) WITH THE NV WEIGHTS, FORMAT (8F10.6), IF IUR .NE. 0
C        CARD WITH INPUT FORMAT, IF IFMT .NE. 0
C           NOTE:  THE FORMAT MUST SPECIFY EXPLICITLY ALL THE SCORES FOR
C           ONE ENTITY, WITH NO IMPLICIT REPITITION IN THE FORMAT, E.G.
C           13 STANDARDIZED SCORES FROM AN SPSS RUN WOULD REQUIRE THE
C           FORMAT (16X,8F8.5/16X,5F8.5).THE DEFAULT FORMAT OF (8F10.6),
C           IF SPECIFIED, IS MODIFIED TO REFLECT THE NUMBER OF SCORES.
C           E.G. FOR 13 SCORES WITH DEFAULT FORMAT, THE FORMAT GENERATED
C           WOULD BE (8F10.6/5F10.6)
C        CARD WITH THE PROBLEM DESCRIPTION IN COL. 1 - 80
C     READ FROM FILE 1...SCORES, ROWWISE, IN SPECIFIED FORMAT
C     WRITE TO FILE 2 THE DISTANCE MATRIX, IN CONTINUOUS (10F10.6)
C     FORMAT
```

```
C     WRITE TO FILE RCPAR THE CARD USED BY EUCDP AND 4 CARDS FOR
C     THE CLUSTER PACKAGE.
C     THIS PROGRAM WAS WRITTEN BY ROBERT B. LEDINGHAM, UNIV. OF WASH.
C
C     READ PARAMETERS AND PUT TOGETHER FORMAT
C
      READ (5, 100) IUR, NH, NV, IFMT, ISQRT, NMETH, SAM, ICS, ISJL,
     1 LBLFLG, RC1
      ISQRT = MINO(MAXO(0, ISQRT), 1)
      IF ( IABS(ICS) .EQ. 0 ) ICS = 0
      J = 1
      IF ( ICS .LT. 0 ) J = 2
      IF ( ICS .GT. 0 ) J = 3
      IF ( NH .GE. 1 ) GO TO 5
      IRD1(1) = IRD2(1)
      IRD1(2) = IRD2(2)
    5 WRITE(6, 120) NH, IRD1, NV, SUS(ISQRT+1), NMETH, RC1, SAM, ICS,
     1 IDCS(1, J), IDCS(2, J), ISJL, LBLFLG
      IF ( IUR .NE. 0 ) READ (5, 101) (WT(I), I = 1, NV)
      IF ( IFMT .EQ. 0 ) GO TO 10
      READ (5, 102) FMT
      GO TO 15
   10 IF ( NV .GE. 9 ) GO TO 11
      FMT(1) = FMT(1) + ISHIFT(NV, 48)
      GO TO 15
   11 NV8 = (NV - 1) / 8
      NVL = NV - NV8 * 8
      IF ( NV8 .GT. 1 ) GO TO 12
      FMT(1) = FMT1(1) + ISHIFT(NVL, 6)
      FMT(2) = FMT1(2)
      GO TO 15
   12 FMT(1) = FMT2(1) + ISHIFT(NV8, 48)
      FMT(2) = FMT2(2) + ISHIFT(NVL, 42)
   15 WRITE(6, 121) FMT
C
C     READ SCORES...IF NH IS SPECIFIED, GET SUFFICIENT MEMORY AND READ.
C     IF NH NOT SPECIFIED, READ A CASE AT A TIME UNTIL EOF.  GET MEMORY
C     AS NEEDED IN 2K BLOCKS, STARTING WITH 1000 WORDS.
C
      IF ( NH .LT. 1 ) GO TO 22
      IW = NV * NH
      IF ( IW .LT. 1001 ) GO TO 20
      MEMORY = ISHIFT(LOCF(X(1)) + IW, 30)
      CALL BAP1(LOCF(MEMORY) + MEMP)
   20 READ (I, FMT) (X(I), I = 1, IW)
      GO TO 29
   22 NH = IWMAX = 0
      IW = 1000
   25 IWMIN = IWMAX + 1
      IWMAX = IWMAX + NV
      NH = NH + 1
      IF ( IWMAX .LT. IW ) GO TO 27
      IW = IW + 2048
      MEMORY = ISHIFT(LOCF(X(1)) + IW, 30)
      CALL BAP1(LOCF(MEMORY) + MEMP)
   27 READ (1, FMT) (X(I), I = IWMIN, IWMAX)
      IF ( EOF(1) ) 29, 25
   29 NH = NH - 1
      IW = IWMIN - 1
      WRITE(6, 126) NH
   29 IMAX = NV
C
C     COMPUTE THE LOWER TRIANGULAR MATRIX...IF NON-UNIT WEIGHTS, USE
C     THEM (CODE THROUGH STMT 40).  IF UNIT WEIGHTS, USE CODE OF STMTS
C     50 THROUGH 70.  IF ISQRT # 0, MATRIX WILL BE UNSQUARED DISTANCES.
C
C     FOR INNER LOOPS COMPUTING ONE ELEMENT OF DIST MAT....IF XP IS
C     ARRAY DIMENSIONED (NV, NH) AND EQUIVALENT TO X ARRAY,
C     X(I), I = IMIN...IMAX CORRESPONDS TO XP(I, N+1), I = 1...NV
C     X(MM), I = IMIN...IMAX CORRESPONDS TO XP(I, M), I = 1...NV
C
      NHM = NH - 1
      K = 0
      IF ( IUR .EQ. 0 ) GO TO 50
```

```
      WRITE(6, 122) (WT(I), I = 1, NV)
      IF ( ISQRT .EQ. 0 ) GO TO 32
      DO 30 I = 1, NV
   30 IF ( WT(I) .LT. 0.0 ) GO TO 31
      GO TO 32
   31 WRITE(6, 128)
   32 CONTINUE
      DO 40 N = 1, NHM
      IMIN = IMAX + 1
      IMAX = IMAX + NV
      MM = 0
      DO 40 M = 1, N
      DK = 0.0
      J = 0
      DO 35 I = IMIN, IMAX
      MM = MM + 1
      J = J + 1
   35 DK = WT(J) * (X(I) - X(MM)) ** 2 + DK
      K = K + 1
      IF ( ISQRT .NE. 0 ) DK = SQRT(DK)
      D(K) = DK
      IF ( K .LT. 10 ) GO TO 40
      WRITE(2, 110) D
      K = 0
   40 CONTINUE
      GO TO 80
   50 WRITE(6, 123)
      DO 70 N = 1, NHM
      IMIN = IMAX + 1
      IMAX = IMAX + NV
      MM = 0
      DO 70 M = 1, N
      DK = 0.0
      DO 60 I = IMIN, IMAX
      MM = MM + 1
   60 DK = (X(I) - X(MM)) ** 2 + DK
      K = K + 1
      IF ( ISQRT .NE. 0 ) DK = SQRT(DK)
      D(K) = DK
      IF ( K .LT. 10 ) GO TO 70
      WRITE(2, 110) D
      K = 0
   70 CONTINUE
   80 IF ( K .GT. 0 ) WRITE(2, 110) (D(I), I = 1, K)
C
C     WRITE RCPAR FILE. WITH FIRST CARD READ BY EUCOP, PROBLEM DESCRIP-
C     TION CARD, TWO CREATED PARAMETER CARDS, AND LABEL/NO LABEL CARD.
C
      WRITE(3, 100) IUR, NH, NV, IFMT, ISQRT, NMETH, SAM, ICS,
     1 ISJL, LBLFLG, RC1
      WRITE(6, 124) IUR, NH, NV, IFMT, ISQRT, NMETH, SAM, ICS,
     1 ISJL, LBLFLG, RC1
      READ (5, 102) FMT
      WRITE(6, 125) FMT
      WRITE(3, 102) FMT
      WRITE(6, 127) NH
      WRITE(3, 103) NH
      IF ( IABS(LBLFLG) .EQ. 0 ) GO TO 90
      WRITE(3, 131)
      WRITE(6, 129)
      STOP
   90 WRITE(3, 132)
      WRITE(6, 130)
      STOP
C
  100 FORMAT(6I4,F4.2,3I4,7A4)
  101 FORMAT(8F10.6)
  102 FORMAT(8A10)
  103 FORMAT(I3,11H 1 0 2  0 2/9H(10F10.6))
  110 FORMAT(10F10.6)
  120 FORMAT(21H-NUMBER OF ENTITIES =,I4,2X,2A10/22HONUMBER OF VARIABLES
     1 =,I3/17HOOUTPUT MATRIX IS,A11,19HEUCLIDEAN DISTANCES/20HONUMBER O
     2F METHODS =,I2,3X,7A6/21HOSAMPLING PARAMETER =,F5.2/25HODENDOGRAM
     3CUTOFF LEVEL =,I3,2X,2A10/23HOS.J.L. OUTPUT OPTION =,I4,24H (NONZE
```

```
     4RO TO WRITE FILE)/22HOENTITY LABEL OPTION =,I4,37H (NONZERO TO REA
     5D LABELS FROM FILE 9))
  121 FORMAT(16HOINPUT FORMAT = ,8A10)
  122 FORMAT(24HOWEIGHTS ARE EXPLICIT.../(1X,8F10.6))
  123 FORMAT(17HOWEIGHTS ARE UNIT)
  124 FORMAT(28HORCPAR FILE (PARAMETERS) .../1HO,6I4,F4.2,3I4,7A4)
  125 FORMAT(1X,8A10)
  126 FORMAT(1HO,I4,14H ENTITIES READ)
  127 FORMAT(1X,I3,11H 1 0 2  0 2/10H (10F10.6))
  128 FORMAT(69H-****  WARNING...YOU HAVE UNSQUARED DISTANCE WITH NEGATI
     1VE WEIGHTS.../38H ****  SOME DISTANCES MAY BE IMAGINARY//)
  129 FORMAT(16H LABELS ON TAPE9)
  130 FORMAT(5H NOLB)
  131 FORMAT(15HLABELS ON TAPE9)
  132 FORMAT(4HNOLB)
      END
*EOR
      PROGRAM RUNCLUS(INPUT = 2038, OUTPUT = 2038, TAPE2, TAPE7, TAPE9,
     1 TAPE5 = INPUT, TAPE6 = OUTPUT)
C
C     PROGRAM RUNCLUS PERFORMS THE FUNCTIONS OF PROGRAM DRIVER MENTIONED
C     IN THE COMMENT SECTION OF SUBROUTINE CNTRL. I.E. IT ASSIGNS I/O
C     UNITS, GETS THE SPACE REQUIRED FOR THE CLUSTERING ALGORITHM, AND
C     CALLS SUBROUTINE CNTRL TO INITIATE THE CLUSTERING PROCESS.
C     THIS VERSION ALLOCATES STORAGE AUTOMATICALLY USING BLANK COMMON.
C     PARAMETER INPUT IS AS FOLLOWS...
C     CARD 1...SAME AS FOR EUCDP CARD 1
C     CARD 2...PROBLEM DESCRIPTION
C     CARD 3...COL 1-3 HAVE NUMBER OF ENTITIES BEING CLUSTERED
C              COL 4-14 HAVE CHARACTERS ' 1 0 2  0 2'
C     CARD 4...FORMAT FOR DISTANCE MATRIX, CURRENTLY (10F10.6)
C     CARD 5...IF COL 1-4 = 4HNOLB, NO LABELS ARE READ
C                         = ANYTHING ELSE, LABELS ARE READ FROM FILE 9
C     IF PROGRAM EUCDP IS USED BEFORE RUNCLUS, EUCDP WILL CONSTRUCT THE
C     PARAMETER FILE FOR RUNCLUS AND SUBSEQUENT PROGRAMS IN THE JOB
C
      COMMON / ICSCOM / ICS
      COMMON X(1)
      DATA MEMP / 4LMEMP /
      READ (5, 100) N, ICS
      IF ( N .GE. 59 ) GO TO 20
      LIMIT = 41 * N
      GO TO 30
   20 LIMIT = N * (N + 25) / 2
   30 MEM = ISHIFT(LOCF(X(1)) + LIMIT, 30)
      WRITE(6, 110) N, LIMIT, MEM
      CALL RAPL(LOCF(MEM) + MEMP)
      CALL CNTRL(X, LIMIT)
      STOP
  100 FORMAT(4X,I4,20X,I4)
  110 FORMAT(21H-NUMBER OF ENTITIES =,I4/27H DYNAMIC STORAGE REQUIRED =,
     1I6,7HD WORDS/31H TOTAL FIELD LENGTH REQUIRED = ,O16,T38,10HB WORDS
     2   )
      END
*EOR
      SUBROUTINE CNTRL(X,LIMIT)
C
C  THIS SUBROUTINE ALLOCATES STORAGE, READS INPUT AND CONTROLS
C  EXECUTION FOR A HIERARCHICAL CLUSTERING JOB BASED ON A PROVIDED
C  SIMILARITY MATRIX.
C
C-----------------------------------------------------------------------
C  INPUT SPECIFICATIONS
C
C  CARD 1  TITLE CARD
C  CARD 2  INFORMATION FOR SUBROUTINES CLSTR AND TREE
C    COLS  1- 3  NE=NUMBER OF ENTITIES (SUBJECTS OR ATTRIBUTES) TO BE
C                   CLUSTERED
C    COLS  4- 5  ISIGN=OPTION FOR SIMILARITY FUNCTION
C                ISIGN=+1, DISTANCE MEASURE
C                ISIGN=-1, CORRELATION MEASURE
C    COLS  6- 7  NTSV=TAPE UNIT ON WHICH CLSTR RESULTS ARE SAVED
C                NTSV=7, PUNCH RESULTS ON CARDS
C                NTSV.LE.0, DO NOT SAVE RESULTS
```

```
C     COLS  8- 9  NTIN=UNIT FROM WHICH SIMILARITY MATRIX IS READ
C                 NTIN=5, CARD READER
C                 NTIN.NE.5, DISK OR TAPE
C     COLS 10-12  INOPT=INPUT OPTION FOR SIMILARITY MATRIX
C                 INOPT.LE.0, THE LOWER TRIANGULAR MATRIX IS STORED AS
C                             ROWS IN ONE LONG LINEAR ARRAY AND READ IN
C                             IN ONE RECORD ON NE*(NE-1)/2 ELEMENTS
C                 INOPT.GT.0, THE LOWER TRIANGULAR MATRIX IS CONSIDERED
C                             TO BE STORED BY ROWS IN ONE LONG LINEAR
C                             ARRAY AND IS READ IN BLOCKS *INOPT* LONG.
C     COLS 13-14  KOUT=OUTPUT OPTION
C                 KOUT=+2, STANDARD OUTPUT
C                 KOUT=-2, STANDARD OUTPUT PLUS PUNCHED SEQUENCE LIST
C                          FROM SUBROUTINE *TREE*
C
C***ANY REPOSITIONING OF THE I/O UNITS NTSV AND NTIN MUST BE
C   ACCOMPLISHED IN PROGRAM DRIVER OR THROUGH USE OF CONTROL CARDS.
C
C   CARD 3  INPUT FORMAT FOR SIMILARITY MATRIX (20A4 FORMAT)
C   CARD(S) 4  SIMILARITY MATRIX
C   CARD 5  END OF RECORD CARD (7/8/9)
C
C***INCLUDE CARDS 4 AND 5 ONLY IF THE SIMILARITY MATRIX IS ON CARDS***
C
C   CARD(S) 6  LABEL CARDS FOR ENTITIES.  THERE ARE TWO OPTIONS
C     1.  INCLUDE 1 CARD WITH THE 4 CHARACTERS *NOLB* IN COLUMNS 1-4.
C         UNDER THIS OPTION LABELS ARE NOT PRINTED ON THE TREE OUTPUT.
C
C     2.  INCLUDE NE CARDS, COLUMNS 1 TO 20 CONTAINING A LABEL FOR ONE
C         ENTITY.  ORDER THE LABEL CARDS IN THE SAME SEQUENCE AS THE
C         ENTITIES ARE REPRESENTED IN THE SIMILARITY MATRIX.
C-----------------------------------------------------------------------
C
C   DECK SETUP SPECIFICATIONS
C
C   THE USER PROVIDES PROGRAM DRIVER WHICH PERFORMS THE FOLLOWING TASKS.
C     1.  ASSIGNS INPUT/OUTPUT UNITS
C     2.  ESTABLISHES THE DIMENSION OF THE X ARRAY AND SETS THIS
C         DIMENSION EQUAL TO LIMIT.
C     3.  CALLS SUBROUTINE CNTRL.
C   THE FOLLOWING EXAMPLE WILL SUFFICE IN MOST CASES.
C
C      PROGRAM DRIVER (INPUT,OUTPUT,PUNCH,TAPE5=INPUT,TAPE6=OUTPUT,
C     1TAPE7=PUNCH,TAPE1,TAPE2)
C      DIMENSION X(7000)
C      LIMIT=7000
C      CALL CNTRL(X,LIMIT)
C      END
C
C   A SECOND JOB DEPENDENT SEGMENT IS SUBROUTINE METHOD.  THE USER
C   SELECTS AMONG THE SEVERAL ALTERNATIVE VERSIONS OF THIS SUBROUTINE TO
C   IMPLEMENT THE DESIRED CLUSTERING TECHNIQUE.
C
C   THE SUBPROGRAMS CNTRL, CLSTR, MTXIN, LFIND AND TREE GO IN EVERY JOB.
C
C   THE X ARRAY IS PARTITIONED FOR STORAGE AS FOLLOWS
C   STORAGE FOR ARRAYS NEEDED AT ALL STAGES OF THE JOB
C   X(N1) TO X(N2-1)  NE WORDS--STORAGE OF THE II ARRAY
C   X(N2) TO X(N3-1)  NE WORDS--STORAGE OF THE JJ ARRAY
C   X(N3) TO X(N4-1)  NE WORDS--STORAGE OF THE SS ARRAY
C   X(N4) TO X(N5-1)  NE WORDS--STORAGE OF THE IL ARRAY
C   X(N5) TO X(N6-1)  NE WORDS--STORAGE OF THE JL ARRAY
C   X(N6) TO X(N7-1)  NE WORDS--STORAGE OF THE NEXT ARRAY
C   STORAGE FOR ARRAYS NEEDED IN SUBROUTINE CLSTR
C   M1=N7
C   X(M1) TO X(M2-1)  (NE*(NE-1))/2 WORDS--STORAGE OF THE S ARRAY
C   X(M2) TO X(M3-1)  NE WORDS--STORAGE OF THE LAST ARRAY
C   X(M3) TO X(M4-1)  NE WORDS--STORAGE OF THE NEAR ARRAY
C   X(M4) TO X(M5-1)  NE WORDS--STORAGE OF THE SREF ARRAY
C   X(M5) TO X(M6-1)  NE WORDS--STORAGE OF THE LIST ARRAY
C   X(M6) TO X(M7-1)  NE WORDS--STORAGE OF THE A ARRAY
C   X(M7) TO X(M8)    NE WORDS--STORAGE OF THE B ARRAY
C   STORAGE FOR ARRAYS NEEDED IN SUBROUTINE TREE (OVERLAY ARRAYS NEEDED
C   IN SUBROUTINE CLSTR)
```

```
C  L1=N7
C  X(L1) TO X(L2-1)  25*NE WORDS--STORAGE OF THE A ARRAY
C  X(L2) TO X(L3-1)  5*NE WORDS--STORAGE OF THE LABEL ARRAY
C  X(L3) TO X(L4-1)  NE WORDS--STORAGE OF THE LCLNO ARRAY
C  X(L4) TO X(L5-1)  NE WORDS--STORAGE OF THE LINE ARRAY
C  X(L5) TO X(L6-1)  NE WORDS--STORAGE OF THE IS ARRAY
C  X(L6) TO X(L7)    NE WORDS--STORAGE OF THE LAST ARRAY
C
      INTEGER FIRST
      DIMENSION X(1),FMT(20),TITLE(20),EPS(25)
      CALL SECOND(TIME)
      WRITE(6,2000) TIME
      READ(5,1000) TITLE
      READ(5,1100) NE,ISIGN,NTSV,NTIN,INOPT,KOUT
      WRITE(6,2500) TITLE
      WRITE(6,2200) NE,ISIGN,NTSV,NTIN,INOPT,KOUT
C  PARTITION THE STORAGE ARRAY
      N1=1
      N2=N1+NE
      N3=N2+NE
      N4=N3+NE
      N5=N4+NE
      N6=N5+NE
      N7=N6+NE
      M2=N7+(NE*(NE-1))/2
      M3=M2+NE
      M4=M3+NE
      M5=M4+NE
      M6=M5+NE
      M7=M6+NE
      M8=M7+NE-1
      L2=N7+25*NE
      L3=L2+6*NE
      L4=L3+NE
      L5=L4+NE
      L6=L5+NE
      L7=L6+NE-1
C  CHECK FOR SUFFICIENT STORAGE
      MAX=M8
      IF(L7.GT.MAX) MAX=L7
      WRITE(6,2300) MAX,LIMIT
      IF(MAX.GT.LIMIT) STOP
C  READ THE SIMILARITY MATRIX
      READ(5,1000) FMT
      WRITE(6,2100) FMT
      CALL MTXIN(X(N7),INOPT,NE,NTIN,FMT)
C  READY TO CLUSTER
60    CALL CLSTR(X(N1),X(N2),X(N3),X(N4),X(N5),X(N6),X(N7),X(M2),X(M3),
     AX(M4),X(M5),X(M6),X(M7),TITLE,NE,ISIGN,NTSV)
C  READ LABEL CARD(S)
      FIRST=L2
      LAST=L2+5
      READ(5,1000) (X(I),I=FIRST,LAST)
      IF(X(FIRST).EQ.4HNOLB) GO TO 80
C  READ REMAINING LABELS
      LAST=L2-1
      DO 70 K=1,NE
      FIRST=LAST+1
      LAST=LAST+6
70    READ(9,1005) (X(I),I=FIRST,LAST)
C  DRAW THE TREE CORRESPONDING TO THE CLUSTERING
80    MERGES=NE-1
      CALL TREE(X(N1),X(N2),X(N3),X(N4),X(N5),X(N6),X(N7),X(L2),X(L3),
     AX(L4),X(L5),X(L6),EPS,TITLE,MERGES,1,6,1,KOUT,NE)
      CALL SECOND (TIME)
      WRITE(6,2000) TIME
      RETURN
1000  FORMAT(20A4)
 1005 FORMAT (//3X,5A5,A4)
1100  FORMAT(I3,3I2,I3,I2)
2000  FORMAT(12H1TIME IS NOW,F10.3,8H SECONDS)
2100  FORMAT(7HOFORMAT,20A4)
2200  FORMAT(5HONE =,I6,/,8H ISIGN =,I5,/,7H NTSV =,I6,/,7H NTIN =,I6,
     A/,8H INOPT =,I5,/,7H KOUT =,I6)
```

```
2300  FORMAT(19HOREQUIRED STORAGE =,I5,6H WORDS,/,
     A       19HOALLOTTED STORAGE =,I5,6H WORDS,/)
2500  FORMAT(1HO,20A4)
      END
*EOR
      FUNCTION LFIND(I,J)
C  IF THE LOWER TRIANGULAR PORTION OF A SYMMETRIC MATRIX IS STORED BY
C  ROWS IN A ONE-DIMENSIONAL ARRAY, THEN THE ELEMENT (I,J) IN THE FULL
C  MATRIX IS ELEMENT LFIND(I,J) IN THE LINEAR ARRAY
      IF(I.GT.J) GO TO 10
C  ROW J, COLUMN I
      LFIND=((J-1)*(J-2))/2+I
      RETURN
C  ROW I, COLUMN J
10    LFIND=((I-1)*(I-2))/2+J
      RETURN
      END
*EOR
      SUBROUTINE CLSTR(II,JJ,SS,IL,JL,NEXT,S,LAST,NEAR,SREF,LIST,A,B,
     ATITLE,N,ISIGN,NT)
C  IN THIS VERSION THE LOWER TRIANGULAR PORTION OF THE SIMILARITY MATRIX
C  IS STORED BY ROWS IN THE ONE-DIMENSIONAL ARRAY S.
C
C  THE FOLLOWING VARIABLES ARE SPECIFIED IN THE CALLING PROGRAM AND
C  ARE PASSED THROUGH THE ARGUMENT LIST
C    N=NUMBER OF OBJECTS TO BE CLUSTERED
C    S(J)=J-TH ELEMENT IN LOWER TRIANGULAR SIMILARITY MATRIX
C    ISIGN=OPTION SPECIFYING TYPE OF SIMILARITY FUNCTION USED
C    ISIGN=+1=DISTANCE MEASURE (DECREASING FUNCTION OF SIMILARITY)
C    ISIGN=-1=CORRELATION MEASURE (INCREASING FUNCTION OF SIMILARITY)
C    NT=TAPE UNIT ON WHICH THE RESULTS ARE SAVED
C    NT.LE.0=DO NOT SAVE RESULTS ON TAPE
C    NT=7=SAVE RESULTS ON PUNCHED CARDS
C    TITLE=IDENTIFYING TITLE FOR THIS RUN
C
C  THE FOLLOWING VARIABLES REPRESENT THE OUTPUT OF THE PROGRAM AND ARE
C  PASSED BACK THROUGH THE ARGUMENT LIST.  THESE RESULTS ARE READY FOR
C  SUBROUTINE TREE.
C    K=STAGE OF CLUSTERING
C    II(K)=LOWER NUMBERED CLUSTER MERGED AT STAGE K
C    JJ(K)=UPPER NUMBERED CLUSTER MERGED AT STAGE K
C    SS(K)=VALUE OF SIMILARITY FUNCTION ASSOCIATED WITH MERGE AT STAGE K
C    IL(K)=PRECEDING STAGE AT WHICH II(K) WAS LAST IN A MERGE
C    JL(K)=PRECEDING STAGE AT WHICH JJ(K) WAS LAST IN A MERGE
C    NEXT(K)=NEXT STAGE AT WHICH II(K) IS IN A MERGE
C
C  IN ADDITION, THE FOLLOWING VARIABLES PLAY IMPORTANT ROLES IN THE PROG
C    NEAR(I)=ID NUMBER OF EXTREME ELEMENT IN ROW I OF THE LOWER
C            TRIANGULAR SIMILARITY MATRIX.
C    SREF(I)=SIMILARITY MEASURE FOR THE PAIR (I,NEAR(I))
C    LIST(I)=I-TH CLUSTER ID NUMBER IN SEQUENTIAL LIST OF CURRENT CLUSTE
C    NCL=NUMBER OF CLUSTERS AT CURRENT STAGE
C    LAST(I)=STAGE NUMBER AT WHICH CLUSTER I WAS LAST IN A MERGE
C    A=WORKING AREA FOR SUBROUTINE METHOD
C    B=WORKING AREA FOR SUBROUTINE METHOD
C
C  THIS SUBROUTINE USES FUNCTION LFIND(I,J) TO FIND THE ADDRESS IN S
C  FOR THE SIMILARITY MEASURE BETWEEN CLUSTERS I AND J
      DIMENSION S(1),II(1),JJ(1),SS(1),IL(1),JL(1),NEXT(1),NEAR(1),
     ASREF(1),LIST(1),LAST(1),A(1),B(1)
      DIMENSION TITLE(20)
C  INITIALIZE VARIABLES AND SET CONSTANTS
      NCL=N
      K=1
      SIGN=ISIGN
      BIG=SIGN*1.E50
      CALL METHOD(S,NEAR,SREF,LIST,A,B,SREFX,SIGN,N,NCL,LREF,NREF,1)
C  INITIALIZE ARRAYS
      DO 10 J=1,N
      LAST(J)=0
      NEXT(J)=0
      LIST(J)=J
      SREF(J)=BIG
10    CONTINUE
```

```
C  FIND EXTREME ENTRY IN EACH ROW
      L=0
      DO 30 I=2,N
      I1=I-1
      DO 30 J=1,I1
      L=L+1
C  IN EFFECT S(L)=S(I,J)
      IF(((S(L)-SREF(I))*SIGN).GT.0.) GO TO 30
      NEAR(I)=J
      SREF(I)=S(L)
30    CONTINUE
C  MAIN LOOP.  FIND EXTREME VALUE IN SREF ARRAY
40    SREFX=BIG
      DO 50 I=2,NCL
      LISTI=LIST(I)
      IF(((SREF(LISTI)-SREFX)*SIGN).GT.0) GO TO 50
      IREF=I
      LREF=LISTI
      SREFX=SREF(LISTI)
50    CONTINUE
C  LREF IS THE ROW NUMBER CONTAINING THE EXTREME ENTRY IN THE S ARRAY.
C  IF THERE ARE TIES, THEN LREF IS THE HIGHEST NUMBERED ROW WITH THIS
C  EXTREME VALUE.  HENCE LREF.GT.NEAR(LREF).  IREF IDENTIFIES THE
C  PLACEMENT OF LREF IN THE LIST ARRAY.
      NREF=NEAR(LREF)
      CALL METHOD(S,NEAR,SREF,LIST,A,B,SREFX,SIGN,N,NCL,LREF,NREF,2)
C  GENERATE MERGE DATA NEEDED FOR SUBROUTINE TREE
      II(K)=NREF
      JJ(K)=LREF
      SS(K)=SREFX
      IL(K)=LAST(NREF)
      JL(K)=LAST(LREF)
      LAST(NREF)=K
      IF(IL(K).EQ.0) GO TO 60
      ILK=IL(K)
      NEXT(ILK)=K
60    IF(JL(K).EQ.0) GO TO 70
      JLK=JL(K)
      NEXT(JLK)=K
70    K=K+1
C  TERMINATE IF N-1 MERGES HAVE OCCURED
      IF(K.EQ.N) GO TO 140
C  UPDATE FOR THE NEXT CYCLE
      NCL=NCL-1
      IF(IREF.GT.NCL) GO TO 90
C  UPDATE LIST ARRAY BY REMOVING LREF AND PUSHING DOWN THE LIST
      DO 80 I=IREF,NCL
80    LIST(I)=LIST(I+1)
C  UPDATE FOR NEXT CYCLE
90    CALL METHOD(S,NEAR,SREF,LIST,A,B,SREFX,SIGN,N,NCL,LREF,NREF,3)
      GO TO 40
C  CLUSTERING FINISHED AND ALL ANCILLARY INFORMATION GENERATED.
C  SAVE RESULTS AS DESIRED.
140   K=K-1
160   IF(NT.LE.0) RETURN
      WRITE(NT,2300) TITLE
      DO 170 I=1,K
170   WRITE(NT,2200) I,II(I),JJ(I),SS(I),IL(I),JL(I),NEXT(I)
      RETURN
2200  FORMAT(3I10,E16.8,3I10)
2300  FORMAT(20A4)
      END
*EOR
      SUBROUTINE MTXIN(X,IOPT,NE,NTIN,FMT)
C  THIS SUBROUTINE READS A LOWER TRIANGULAR MATRIX *X* REPRESENTING
C  ASSOCIATION AMONG *NE* ENTITIES.  THE MATRIX IS READ FROM UNIT *NTIN*
C  IN FORMAT *FMT*.  THE MODE OF INPUT FOR THE MATRIX IS DETERMINED BY
C  THE *IOPT* PARAMETER AS FOLLOWS.
C    IOPT.LE.0, MATRIX IS READ IN LOWER TRIANGULAR FORM BY ROWS, EACH
C               ROW BEING A NEW RECORD.
C    IOPT.GT.0, MATRIX IS READ IN CONSTANT LENGTH BLOCKS, EACH *IOPT*
C               WORDS LONG.
      DIMENSION FMT(20),X(1)
      INTEGER FIRST
```

```
      IF(IOPT.LE.0) GO TO 30
C  READ THE SIMILARITY MATRIX IN BLOCKS *IOPT* LONG
      FIRST=1
      LAST=IOPT
10    READ(NTIN,FMT) (X(I),I=FIRST,LAST)
C  USE THE END OF RECORD CARD TO SIGNIFY END OF THE SIMILARITY MATRIX
      IF( EOF(NTIN) ) 60,20
20    FIRST=FIRST+IOPT
      LAST=LAST+IOPT
      GO TO 10
C  READ THE SIMILARITY MATRIX AS ROWS OF A LOWER TRIANGULAR MATRIX,
C  IN ONE RECORD OF NC WORDS
   30 NC=(NE**2-NE)/2
      READ(NTIN,FMT) (X(I),I=1,NC)
      IF ( EOF(NTIN) ) 200,40
C  PASS THE END OF FILE
   40 READ(NTIN,FMT) Z
      IF( EOF(NTIN) ) 60,210
60    RETURN
C  ERROR MESSAGES
200   WRITE(6,2400)
      STOP
210   WRITE(6,2500)
      STOP
2400  FORMAT(36HOEOF ENCOUNTERED WHEN NONE EXPECTED.)
2500  FORMAT(30HONO EOF WHEN ONE WAS EXPECTED.)
      END
*EOR
      SUBROUTINE TREE(I,J,S,IL,JL,NEXT,A,LABEL,LCLNO,LINE,IS,LAST,EPS,
     ATITLE,N,KBEG,NT,INTRV,IPRNT,MAXIN)
C
C  DATA INPUT THROUGH CALLING SEQUENCE
C
C  N=HIGHEST STAGE NUMBER IN THE CLUSTER MERGE DATA (MUST BE EXACT)
C  KBEG=STAGE NUMBER AT WHICH THE TREE BEGINS, DEFAULT VALUE 1
C  NT=TAPE NUMBER FOR PRINTED OUTPUT, DEFAULT VALUE = 6
C  INTRV=INTERVAL OPTION FOR SEGMENTATION
C  INTRV=1=DEFAULT VALUE, CONSTRUCT EPS BY DIVIDING THE RANGE OF S INTO
C          25 EQUAL SEGMENTS
C  INTRV=2=EPS IS PROVIDED AS PART OF THE ARGUMENT LIST
C  INTRV=3=THE IS ARRAY IS ALREADY CONSTRUCTED AND EPS IS PROVIDED FOR T
C  IPRNT=PRINT OPTION FOR INPUT INFORMATION
C  IABS(IPRNT)=1, PRINT ONLY TITLE AND *IS* ARRAY
C  IABS(IPRNT).NE.1, IN ADDITION PRINT THE CLUSTER MERGE DATA
C  IPRNT.LE.0, IN ADDITION, PUNCH THE SEQUENCE IN WHICH THE ENTITIES
C              APPEAR IN THE TREE (NEEDED FOR POST-ANALYSIS OF DATA
C              UNIT CLUSTERING IN SUBROUTINE *POSTDU*).
C  EPS(M)=RIGHT ENDPOINT FOR THE MTH INTERVAL USED FOR SEGMENTING S
C  LABEL(M,IJ)=MTH OF 5 WORDS IDENTIFYING THE IJTH OBJECT
C  TITLE=ARRAY OF 20 WORDS FOR IDENTIFYING THE RUN.
C  K=INDEX IDENTIFYING STAGE NUMBER IN THE CLUSTERING
C  I(K)=LOWER NUMBERED CLUSTER IDENTIFICATION NUMBER IN THE MERGE AT THE
C         KTH STAGE
C  J(K)=UPPER NUMBERED CLUSTER IDENTIFICATION NUMBER IN THE MERGE AT THE
C         KTH STAGE
C  S(K)=VALUE OF THE CRITERION FUNCTION FOR THE MERGE AT THE KTH STAGE
C  IS(K)=CATEGORIZED VALUE OF S= INTEGER IN RANGE 1 TO 25
C  IL(K)=STAGE NUMBER WHEN I(K) WAS LAST IN A MERGE (0 FOR FIRST MERGE F
C  JL(K)=STAGE NUMBER WHEN J(K) WAS LAST IN A MERGE (0 FOR FIRST MERGE F
C  NEXT(K)=STAGE NUMBER WHEN I(K) IS NEXT IN A MERGE
C  MAXIN=HIGHEST CLUSTER ID NUMBER IN THE CLUSTER MERGE DATA
C
C  OTHER VARIABLES USED IN THE PROGRAM
C
C  LINE(I)=LINE NUMBER IN THE PRINTOUT AT WHICH I(K) IS CARRIED (AFTER
C          MOST RECENT MERGE)
C  LCLNO(L)=THE CLUSTER NUMBER TO BE PRINTED ON LINE L AT THE LEFT OF TH
C  A(M,L)=THE MTH SEGMENT (OF 25) IN THE LTH LINE OF THE PRINTOUT
C  LAST(L)=FARTHEST RIGHT SEGMENT IN LINE L WHICH IS NOT BLANK
C
C     IN ADDITION, COMMON BLOCK /ICSCOM/ PROVIDES PARAMETER ICS TO ALLOW
C     ONLY A PORTION (HORIZONTALLY) OF THE TREE TO BE SHOWN, THUS EXPAND
C     ING THE DETAIL OF THE TREE, SINCE AN EXTREME ISOLET CAN HAVE THE
C     EFFECT OF SQUASHING MERGES TOGETHER IN THE REST OF THE TREE.
```

```
C      DEFAULT ICS = 0 PRINTS THE ENTIRE TREE.  ICS > 0 PRINTS MERGES OF
C      LEVEL UP THROUGH THAT OF THE ICS-TH MERGE.  ICS < 0 SELECTS AUTO-
C      CUTOFF...LOOKING AT THE LAST 5 MERGES IN REVERSE ORDER, IF THE
C      MERGE EXISTS UNIQUELY FOR MORE THAN 1/5 THE ENTIRE RANGE OF DIST-
C      ANCES, THE MAXIMUM DISTANCE SCALED IN THE DENDOGRAM IS SET
C      TO THE DISTANCE OF THE NEXT HIGHEST MERGE.
C
       DIMENSION I(N),J(N),S(N),IS(N),IL(N),JL(N),NEXT(N),LCLNO(N),
      AA(25,N),LAST(N)
       DIMENSION LINE(MAXIN),LABEL(6,MAXIN)
       DIMENSION EPS(25),TITLE(20)
       COMMON / ICSCOM / ICS
       REAL LABEL
       DATA BAR1,BLNKT,BARS,BLANK/4H---I,4H    I,4H----,4H     /
       DATA ICS / 0 /
C  DEFAULT VALUES
       KBEG = MAXO(KBEG, 1)
       IF(INTRV.LT.1.OR.INTRV.GT.3) INTRV=1
       IF(NT.LE.0) NT=6
C  INITIALIZE ARRAYS
       NOBJ=N+1
       DO 10 K=1,NOBJ
       LINE(K)=0
       LCLNO(K)=0
       LAST(K)=0
       DO 10 L=1,25
       A(L,K)=BLANK
10     CONTINUE
C  SEGMENT THE S ARRAY
       IF ( INTRV - 2 ) 20, 40, 120
C  CONSTRUCT INTERVALS OF EQUAL LENGTH
   20  SRMAX = S(N)
       IF ( ICS ) 24, 26, 22
   22  SRMAX = S(ICS)
       GO TO 28
C    ICS < 0...USE AUTO-CUTOFF
   24  DO 26 L = 1, 5
       IF ( S(N-L+1) - S(N-L) .GT. (SRMAX - S(1)) / 5.0 ) SRMAX = S(N-L)
   26  CONTINUE
   28  RANGE = SRMAX - S(KBEG)
       DELTA=RANGE/25.
       EPS(1)=S(KBEG)+DELTA
       DO 30 K=2,24
30     EPS(K)=EPS(K-1)+DELTA
       EPS(25)=S(N)
C  CONSTRUCT THE IS ARRAY
40     IF(EPS(1).GT.EPS(2)) GO TO 70
C  S INCREASES WITH DISSIMILARITY (AS DOES A DISTANCE)
       KK=1
       DO 60 K=1,N
50     IF(S(K).LE.EPS(KK)) GO TO 60
       IF(KK.EQ.25) GO TO 60
       KK=KK+1
       GO TO 50
60     IS(K)=KK
       GO TO 120
C  S DECREASES WITH DISSIMILARITY (AS DOES A CORRELATION)
70     KK=24
       KKK=25
       NN=N+1
       DO 90 K=1,N
       KCOMP=NN-K
80     IF(S(KCOMP).LT.EPS(KK)) GO TO 90
       KKK=KK
       KK=KK-1
       IF(KK.EQ.0) GO TO 100
       GO TO 80
90     IS(KCOMP)=KKK
100    DO 110 K=1,KCOMP
110    IS(K)=1
C  PRINT INPUT TO TREE
120    WRITE(NT,2000) TITLE
       WRITE(NT,2100) KBEG,N
       WRITE(NT,2200)
```

```
      WRITE(NT,2300)
      M=1
      WRITE(NT,2400) M,S(KBEG),EPS(M)
      DO 130 M=2,25
      MM=M-1
130   WRITE(NT,2400) M,EPS(MM),EPS(M)
      IF(IABS(IPRNT).EQ.1) GO TO 150
C  PRINT THE CLUSTER MERGE DATA
      WRITE(NT,2000)  TITLE
      WRITE(NT,2500)
      DO 140 K=KBEG,N
      WRITE(NT,2600) K,I(K),J(K),S(K),IS(K),IL(K),JL(K),NEXT(K)
      WRITE(7,4321) K,I(K),J(K),S(K),IL(K),JL(K),IS(K)
 4321 FORMAT (3I5,E16.8,3I5)
140   CONTINUE
C  START TREE WITH THE MOST SIMILAR PAIR
150   K=KBEG
      LNO=0
C  MERGE CLUSTERS I(K) AND J(K)
160   IK=I(K)
      JK=J(K)
C  SET LINE NUMBERS FOR OUTPUT
      IF(IL(K).NE.0) GO TO 170
      LNO=LNO+1
      LINE(IK)=LNO
      LCLNO(LNO)=IK
170   IF(JL(K).NE.0) GO TO 180
      LNO=LNO+1
      LINE(JK)=LNO
      LCLNO(LNO)=JK
C  FILL IN THE PRINT LINES
180   ISK=IS(K)
      KT=0
      ITEM=IK
190   LITEM=LINE(ITEM)
      IF(ISK-LAST(LITEM)-1) 225,200,210
C  ADD ONLY ONE MORE SEGMENT FOR LINE(ITEM)
200   A(ISK,LITEM)=BARI
      LAST(LITEM)=ISK
      GO TO 225
C  ADD MORE THAN ONE SEGMENT
210   LBEG=LAST(LITEM)+1
      LEND=ISK-1
      DO 220 L=LBEG,LEND
220   A(L,LITEM)=BARS
      GO TO 200
C  REPEAT FOR CLUSTER J(K)
225   KT=KT+1
      IF(KT.NE.1) GO TO 230
      ITEM=JK
      GO TO 190
C  TAKE CARE OF ANY LINES BETWEEN I(K) AND J(K)
230   LIK=LINE(IK)
      LJK=LINE(JK)
      IF(LIK.GT.LJK) GO TO 240
      LBOT=LJK
      LTOP=LIK
      GO TO 250
240   LBOT=LIK
      LTOP=LJK
250   IF(LBOT.EQ.(LTOP+1)) GO TO 270
C  MUST FILL IN SOME VERTICAL CONNECTIONS
      LBEG=LTOP+1
      LEND=LBOT-1
      DO 260 L=LBEG,LEND
      IF(A(ISK,L).EQ.BARI) GO TO 260
      A(ISK,L)=BLNKI
      LAST(L)=ISK
260   CONTINUE
C  UPDATE LINE NUMBER FOR NEW CLUSTER
270   LINE(IK)=(LINE(IK)+LINE(JK))/2
C  MERGE COMPLETE.  FIND NEXT STAGE
      KLAST=K
      K=NEXT(K)
```

```
      IF(K.GT.N.OR.K.LT.KREG) GO TO 400
      IF(IL(K).LE.0) GO TO 290
      IF(JL(K).LE.0) GO TO 290
      GO TO 300
280   IL(K)=-IL(K)
      GO TO 160
290   JL(K)=-JL(K)
      GO TO 160
C  THIS MERGE INVOLVES CLUSTERS THAT EACH HAVE MORE THAN ONE MEMBER.
C  BACKTRACK TO THE ROOT OF THE TREE ALONG THE UNEXPLORED BRANCH.
300   IF(IL(K).EQ.KLAST) GO TO 310
C  GO DOWN IL(K) BRANCH.  SET JL(K) SO WE KNOW NOT TO GO DOWN THAT BRANC
      JL(K)=-JL(K)
      K=IL(K)
      GO TO 320
C  GO DOWN JL(K) BRANCH.  SET IL(K) SO WE KNOW NOT TO GO DOWN THAT BRANC
310   IL(K)=-IL(K)
      K=JL(K)
320   IF(K.LT.1.OR.K.GT.N) GO TO 600
C  TEST TO SEE IF THE END HAS BEEN REACHED.  IL(K)=JL(K) IFF BOTH ZERO.
      IF(IL(K)-JL(K)) 330,160,350
330   IF(IL(K).EQ.0) GO TO 360
340   K=IL(K)
      GO TO 320
350   IF(JL(K).EQ.0) GO TO 340
360   K=JL(K)
      GO TO 320
C  PRINT THE TREE
400   WRITE(NT,2000) TITLE
      WRITE(NT,3000) (K,K=1,25)
      ENDFILE 7
      IF(LABEL(1,1).EQ.4HNOLB) GO TO 420
      DO 410 L=1,LNO
      LL=LCLNO(L)
      WRITE(7,417) LL
  417 FORMAT (I5)
410   WRITE(NT,3100) (LABEL(K,LL),K=1,6),LL,(A(K,L),K=1,25)
      GO TO 440
C  LEAVE LABEL SPACES BLANK
420   DO 430 L=1,LNO
      LL=LCLNO(L)
      WRITE(7,417) LL
430   WRITE(NT,3200) LL,(A(K,L),K=1,25)
C  TREE COMPLETE
440   WRITE(NT,3000) (K,K=1,25)
      ENDFILE NT
      IF(IPRNT.GT.0) RETURN
C  PUNCH SEQUENCE LIST
      WRITE(7,3900) TITLE
      WRITE(7,4000) (LCLNO(L),L=1,LNO)
      RETURN
C  ERROR.  PRINT AS MUCH OF THE TREE AS HAS BEEN CONSTRUCTED
600   WRITE(NT,6000) KLAST,K
      GO TO 400
2000  FORMAT(1H1,20X,20A4)
2100  FORMAT(65H0THIS RUN DEPICTS THE PORTION OF THE TREE GENERATED BETW
     AEEN STAGE, I5,10H AND STAGE,,I5,19H OF THE CLUSTERING.)
2200  FORMAT(63H0THE CRITERION VALUES ARE SEGMENTED INTO THE FOLLOWING C
     ALASSES.)
2300  FORMAT(6H0CLASS,5X,11HLOWER BOUND,5X,11HUPPER BOUND)
2400  FORMAT(1X,I5,2E16.8)
2500  FORMAT(1H0,9X,1HK,9X,1HI,9X,1HJ,15X,1HS,8X,2HIS,8X,2HIL,8X,2HJL,6X
     A,4HNEXT)
2600  FORMAT(1X,3I10,E16.8,4I10)
3000  FORMAT(10H0ITEM NAME,19X,5HID NO,1X,25I4)
C  IF LOCAL CONVENTIONS PERMIT, RECOMMEND THAT THE CARRIAGE CONTROL
C  CHARACTER IN FORMATS 3100 AND 3200 ALLOW 66 LINES OF PRINT PER PAGE.
C  THAT IS, THE MARGINS AT THE TOP AND BOTTOM OF THE PAGE ARE SUPPRESSED
C  AND PRINTING IS SINGLE SPACE.
3100  FORMAT (1H ,5A5,A4,1X,I4,25A4)
3200  FORMAT(30X,I5,25A4)
3900  FORMAT(20A4)
4000  FORMAT(20I4)
6000  FORMAT(37H0ERROR. WHILE BACKTRACKING FROM KLAST,I6,26H K WAS FOUND
```

```
     A OUT OF RANGE.,/,1X,3HK =,I20)
      END
*EOR
      SUBROUTINE METHOD(S,NEAR,SREF,LIST,A,B,SREFX,SIGN,N,NCL,LREF,NREF,
     AJOB)
C
C  HIERARCHICAL CLUSTERING BY SINGLE LINKAGE.  ALGORITHM IS DERIVED
C  FROM
C  JOHNSON, S.C., HIERARCHICAL CLUSTERING SCHEMES, PSYCHOMETRIKA,
C  VOLUME 32, NUMBER 3, SEPTEMBER 1967, PP241-254.
C
      DIMENSION S(1),NEAR(1),SREF(1),LIST(1),A(1),B(1)
      IF ( JOB - 2 ) 10, 15, 20
C  JOB=1, INITIALIZATION
   10 WRITE(6,2000)
 2000 FORMAT(26HOSINGLE LINKAGE CLUSTERING)
      BIG = SIGN*1.E50
      RETURN
C  JOB=2, DUMMY ENTRY.
   15 RETURN
C  JOB=3, UPDATE FOR NEXT ROUND.
   20 SREF(NREF)=BIG
      DO 50 J=1,NCL
C  UPDATE ENTRIES IN S ARRAY ASSOCIATED WITH NREF
      I=LIST(J)
      IF(I.EQ.NREF) GO TO 50
C  RECALL THAT LREF HAS BEEN REMOVED FROM LIST SO I NEED NOT BE
C  TESTED FOR EQUALITY WITH LREF
      LL=LFIND(I,LREF)
      LN=LFIND(I,NREF)
      IF(((S(LL)-S(LN))*SIGN).GE.0.) GO TO 30
      S(LN)=S(LL)
   30 IF(I.GT.NREF) GO TO 40
C  CHECK WHETHER S(LN) IS A BETTER CANDIDATE FOR SREF(NREF)
      IF(((S(LN)-SREF(NREF))*SIGN).GT.0.) GO TO 50
      NEAR(NREF)=I
      SREF(NREF)=S(LN)
      GO TO 50
C  UPDATE NEAR ARRAY FOR THOSE ROWS WHOSE EXTREME ELEMENT WAS LREF
   40 IF(NEAR(I).NE.LREF.AND.NEAR(I).NE.NREF) GO TO 50
      NEAR(I)=NREF
      SREF(I)=S(LN)
   50 CONTINUE
      RETURN
      END
*EOR
      SUBROUTINE METHOD(S,NEAR,SREF,LIST,A,B,SREFX,SIGN,N,NCL,LREF,NREF,
     AJOB)
C
C  HIERARCHICAL CLUSTERING BY COMPLETE LINKAGE.  THE ALGORITHM IS
C  DERIVED FROM
C  JOHNSON, S.C., HIERARCHICAL CLUSTERING SCHEMES, PSYCHOMETRIKA,
C  VOLUME 32, NUMBER 3, SEPTEMBER 1967, PP 241-254.
C
      DIMENSION S(1),NEAR(1),SREF(1),LIST(1),A(1),B(1)
      IF ( JOB - 2 ) 10, 15, 20
C  JOB=1, INITIALIZATION
10    WRITE(6,2000)
2000  FORMAT(28HOCOMPLETE LINKAGE CLUSTERING)
      BIG=SIGN*1.E50
      RETURN
C  JOB=2, DUMMY ENTRY.
15    RETURN
C  JOB=3, UPDATE FOR NEXT ROUND.
20    DO 30 J=1,NCL
      I=LIST(J)
      IF(I.EQ.NREF) GO TO 30
C  RECALL THAT LREF HAS BEEN REMOVED FROM LIST SO I NEED NOT BE
C  TESTED FOR EQUALITY WITH LREF.
      LL=LFIND(I,LREF)
      LN=LFIND(I,NREF)
      IF(((S(LL)-S(LN))*SIGN).LE.0) GO TO 30
      S(LN)=S(LL)
30    CONTINUE
```

```
C  UPDATE THE NEAR AND SREF ARRAYS.  IF THE EXTREME ELEMENT IN ROW I
C  WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW EXTREME
C  ELEMENT.  ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
40     DO 50 J=1,NCL
       I=LIST(J)
       IF(I .EQ. NREF) GO TO 55
50     CONTINUE
55     IF(J.EQ.1) GO TO 80
60     SREF(I)=BIG
       J1=J-1
       DO 70 L=1,J1
       LISTL=LIST(L)
       LL=LFIND(I,LISTL)
       IF(((S(LL)-SREF(I))*SIGN).GE.0.) GO TO 70
       NEAR(I)=LISTL
       SREF(I)=S(LL)
70     CONTINUE
80     J=J+1
       IF(J.GT.NCL) RETURN
       I=LIST(J)
       IF(NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60
       GO TO 80
       END
*EOR
      SUBROUTINE METHOD(S,NEAR,SREF,LIST,NUMBR,SUM,SREFX,SIGN,N,NCL,
     ALREF,NREF,JOB)
C
C  HIERARCHICAL CLUSTERING BY MINIMIZING THE AVERAGE DISTANCE OR
C  MAXIMIZING THE AVERAGE CORRELATION BETWEEN THE MERGED GROUPS.
C
C  THE ALGORITHM IS DERIVED FROM THE *GROUP AVERAGE* METHOD DESCRIBED IN
C    LANCE, G.N. AND W.T. WILLIAMS, A GENERAL THEORY OF CLASSIFICATORY
C    SORTING STRATEGIES, 1. HIERARCHICAL SYSTEMS, THE COMPUTER JOURNAL,
C    VOLUME 9, NUMBER 4, FEBRUARY 1967, PP373-380.
C
       DIMENSION S(1),NEAR(1),SREF(1),LIST(1),NUMBR(1),SUM(1)
       IF ( JOB - 2 ) 10, 25, 30
C  JOB=1, INITIALIZE.
C  NUMBR(I)=NUMBER OF ENTITIES CURRENTLY IN THE I-TH CLUSTER
10     WRITE(6,2000)
2000   FORMAT(42HOAVERAGE LINKAGE BETWEEN THE MERGED GROUPS)
       DO 20 J=1,N
20     NUMBR(J)=1
       BIG=SIGN*1.E50
       RETURN
C  JOB=2, DUMMY ENTRY.
25     RETURN
C  JOB=3, UPDATE FOR NEXT ROUND.
C  UPDATE THE NEW CLUSTER
30     NUMBR(NREF)=NUMBR(NREF)+NUMBR(LREF)
C  UPDATE ENTRIES IN THE REDUCED SIMILARITY MATRIX.  THE ENTRIES ARE
C  THE SUM TOTAL OF SIMILARITY VALUES ASSOCIATED WITH ALL
C  PAIRWISE LINKS BETWEEN THE ELEMENTS OF THE TWO CLUSTERS.
       DO 40 J=1,NCL
       I=LIST(J)
       IF(I.EQ.NREF) GO TO 40
C  RECALL THAT LREF HAS BEEN REMOVED FROM LIST AND THEREFORE I NEED NOT
C  BE TESTED FOR EQUALITY WITH LREF.
       LL=LFIND(I,LREF)
       LN=LFIND(I,NREF)
       S(LN)=S(LN)+S(LL)
40     CONTINUE
C  WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW EXTREME
C  ELEMENT.  ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
       DO 50 J=1,NCL
       I=LIST(J)
       IF(I.EQ.NREF) GO TO 55
50     CONTINUE
55     IF(J.EQ.1) GO TO 80
60     SREF(I)=BIG
       J1=J-1
       DO 70 L=1,J1
       LISTL=LIST(L)
       LL=LFIND(I,LISTL)
```

```
      SREFX=S(LL)/(NUMBR(I)*NUMBR(LISTL))
      IF(((SREFX-SREF(I))*SIGN).GE.0.) GO TO 70
      NEAR(I)=LISTL
      SREF(I)=SREFX
70    CONTINUE
80    J=J+1
      IF(J.GT.NCL) RETURN
      I=LIST(J)
      IF(NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60
      GO TO 90
      END
*EOR
      SUBROUTINE METHOD(S,NEAR,SREF,LIST,NUMBR,SUM,SREFX,SIGN,N,NCL,
     ALREF,NREF,JOB)
C
C  HIERARCHICAL CLUSTERING BY MINIMIZING THE AVERAGE DISTANCE OR
C  MAXIMIZING THE AVERAGE CORRELATION WITHIN THE NEW GROUP.  THAT IS,
C  FOR EACH POTENTIAL MERGE THE AVERAGE OF ALL LINKAGES WITHIN THE
C  NEW GROUP IS CALCULATED.
      DIMENSION S(1),NEAR(1),SREF(1),LIST(1),NUMBR(1),SUM(1)
      IF ( JOB - 2 ) 10, 25, 30
C  JOB=1, INITIALIZE.
C  NUMBR(I)=NUMBER OF ENTITIES CURRENTLY IN THE I-TH CLUSTER
C  SUM(I)=SUM OF ALL PAIRWISE SIMILARITIES AMONG ENTITIES IN THE I-TH
C         CLUSTER
10    WRITE(6,2000)
2000  FORMAT(37HOAVERAGE LINKAGE WITHIN THE NEW GROUP)
      DO 20 J=1,N
      NUMBR(J)=1
20    SUM(J)=0.
      BIG=SIGN*1.E50
      RETURN
C  JOB=2, DUMMY ENTRY.
25    RETURN
C  JOB=3, UPDATE FOR NEXT ROUND.
C  UPDATE THE NEW CLUSTER
30    NUMBR(NREF)=NUMBR(NREF)+NUMBR(LREF)
      LN=LFIND(LREF,NREF)
      SUM(NREF)=SUM(NREF)+SUM(LREF)+S(LN)
C  UPDATE ENTRIES IN THE REDUCED SIMILARITY MATRIX.  THE ENTRIES ARE
C  THE SUM TOTAL OF SIMILARITY VALUES ASSOCIATED WITH ALL
C  PAIRWISE LINKS BETWEEN THE ELEMENTS OF THE TWO CLUSTERS.
      DO 40 J=1,NCL
      I=LIST(J)
      IF(I.EQ.NREF) GO TO 40
C  RECALL THAT LREF HAS BEEN REMOVED FROM LIST AND THEREFORE I NEED NOT
C  BE TESTED FOR EQUALITY WITH LREF.
      LL=LFIND(I,LREF)
      LN=LFIND(I,NREF)
      S(LN)=S(LN)+S(LL)
40    CONTINUE
C  UPDATE THE NEAR AND SREF ARRAYS.  IF THE EXTREME ELEMENT IN ROW I
C  WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW EXTREME
C  ELEMENT.  ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
      DO 50 J=1,NCL
      I=LIST(J)
      IF(I.EQ.NREF) GO TO 55
50    CONTINUE
55    IF(J.EQ.1) GO TO 80
60    SREF(I)=BIG
      J1=J-1
      DO 70 L=1,J1
      LISTL=LIST(L)
      LL=LFIND(I,LISTL)
      NTOT=NUMBR(I)+NUMBR(LISTL)
      NTOT=(NTOT*(NTOT-1))/2
      SREFX=(SUM(I)+SUM(LISTL)+S(LL))/NTOT
      IF(((SREFX-SREF(I))*SIGN).GE.0.) GO TO 70
      NEAR(I)=LISTL
      SREF(I)=SREFX
70    CONTINUE
80    J=J+1
      IF(J.GT.NCL) RETURN
      I=LIST(J)
```

```
      IF(NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60
      GO TO 80
      END
*EOR
      SUBROUTINE METHOD(S,NEAR,SREF,LIST,A,B,SREFX,SIGN,N,NCL,LREF,NREF,
     AJOB)
C
C  HIERARCHICAL CLUSTERING BY THE MEDIAN METHOD OF
C    GOWER, J.C., A COMPARISON OF SOME METHODS OF CLUSTER ANALYSIS,
C    BIOMETRICS, VOLUME 23, NUMBER 4, DECEMBER 1967, PP 623-637.
C
C  THE PARTICULAR ALGORITHM USED HERE IS DESCRIBED IN
C    LANCE, G.N. AND W.T. WILLIAMS, A GENERAL THEORY OF CLASSIFICATORY
C    SORTING STRATEGIES, 1. HIERARCHICAL SYSTEMS, THE COMPUTER JOURNAL,
C    VOLUME 9, NUMBER 4, FEBRUARY 1967. PP373-380.
C
      DIMENSION S(1),NEAR(1),SREF(1),LIST(1),A(1),B(1)
      IF ( JOB - 2 ) 10, 15, 20
C  JOB=1. INITIALIZATION
10    WRITE(6,2000)
2000  FORMAT(44HOMEDIAN METHOD OF GOWER, BEWARE OF REVERSAL(S)
      BIG=SIGN*1.E50
      RETURN
C  JOB=2, DUMMY ENTRY.
15    RETURN
C  JOB=3, UPDATE FOR NEXT ROUND.
20    LRET=LFIND(LREF,NREF)
      DO 30 J=1,NCL
      I=LIST(J)
      IF(I.EQ.NREF) GO TO 30
C  RECALL THAT LREF HAS BEEN REMOVED FROM LIST SO I NEED NOT BE
C  TESTED FOR EQUALITY WITH LREF.
      LL=LFIND(I,LREF)
      LN=LFIND(I,NREF)
      S(LN)=(S(LN)+S(LL))/2.-S(LRET)/4.
30    CONTINUE
C  UPDATE THE NEAR AND SREF ARRAYS.  IF THE EXTREME ELEMENT IN ROW I
C  WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW EXTREME
C  ELEMENT.  ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
40    DO 50 J=1,NCL
      I=LIST(J)
      IF(I.EQ.NREF) GO TO 55
50    CONTINUE
55    IF(J.EQ.1) GO TO 80
60    SREF(I)=BIG
      J1=J-1
      DO 70 L=1,J1
      LISTL=LIST(L)
      LL=LFIND(I,LISTL)
      IF(((S(LL)-SREF(I))*SIGN).GE.0.) GO TO 70
      NEAR(I)=LISTL
      SREF(I)=S(LL)
70    CONTINUE
80    J=J+1
      IF(J.GT.NCL) RETURN
      I=LIST(J)
      IF(NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60
      GO TO 80
      END
*EOR
      SUBROUTINE METHOD(S,NEAR,SREF,LIST,NUMBR,SUM,SREFX,SIGN,N,NCL,
     ALREF,NREF,JOB)
C
C  HIERARCHICAL CLUSTERING BY CENTROID SORTING
C
C  THE PARTICULAR ALGORITHM USED HERE IS DESCRIBED IN
C    LANCE, G.N. AND W.T. WILLIAMS, A GENERAL THEORY OF CLASSIFICATORY
C    SORTING STRATEGIES, 1. HIERARCHICAL SYSTEMS, THE COMPUTER JOURNAL,
C    VOLUME 9, NUMBER 4, FEBRUARY 1967, PP373-380.
C
      DIMENSION S(1),NEAR(1),SREF(1),LIST(1),NUMBR(1),SUM(1)
      IF ( JOB - 2 ) 10, 25, 30
C  JOB=1, INITIALIZE.
C  NUMBR(I)=NUMBER OF ENTITIES CURRENTLY IN THE I-TH CLUSTER
```

```
C          CLUSTER
10     WRITE(6,2000)
2000   FORMAT(42HOCENTROID CLUSTERING.  BEWARE OF REVERSALS)
       DO 20 J=1,N
20     NUMBR(J)=1
       BIG=SIGN*1.E50
       RETURN
C  JOB=2, DUMMY ENTRY.
25     RETURN
C  JOB=3, UPDATE FOR NEXT ROUND.
C  UPDATE THE NEW CLUSTER
30     NTOT=NUMBR(NREF)+NUMBR(LREF)
       TOT=NTOT
       ALL=NUMBR(LREF)/TOT
       ALN=NUMBR(NREF)/TOT
       NUMBR(NREF)=NTOT
       PROD=ALN*ALL
       LRET=LFIND(LREF,NREF)
       DO 40 J=1,NCL
       I=LIST(J)
       IF(I.EQ.NREF) GO TO 40
C  RECALL THAT LREF HAS BEEN REMOVED FROM LIST AND THEREFORE I NEED NOT
C  BE TESTED FOR EQUALITY WITH LREF.
       LL=LFIND(I,LREF)
       LN=LFIND(I,NREF)
       S(LN)=ALL*S(LL)+ALN*S(LN)-PROD*S(LRET)
40     CONTINUE
C  UPDATE THE NEAR AND SREF ARRAYS.  IF THE EXTREME ELEMENT IN ROW I
C  WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW EXTREME
C  ELEMENT.  ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
       DO 50 J=1,NCL
       I=LIST(J)
       IF(I.EQ.NREF) GO TO 55
50     CONTINUE
55     IF(J.EQ.1) GO TO 80
60     SREF(I)=BIG
       J1=J-1
       DO 70 L=1,J1
       LISTL=LIST(L)
       LI=LFIND(I,LISTL)
       IF(((S(LL)-SREF(I))*SIGN).GE.0.) GO TO 70
       NEAR(I)=LISTL
       SREF(I)=S(LL)
70     CONTINUE
80     J=J+1
       IF(J.GT.NCL) RETURN
       I=LIST(J)
       IF(NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60
       GO TO 80
       END
*EOR
       SUBROUTINE METHOD(S,NEAR,SREF,LIST,NUMBR,SUM,SREFX,SIGN,N,NCL,
      ALREF,NREF,JOB)
C
C  HIERARCHICAL CLUSTERING BY THE METHOD OF
C    WARD, J.H., JR, HIERARCHICAL GROUPING TO OPTIMISE AN OBJECTIVE
C    FUNCTION, JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, VOLUME
C    58, 1963, PP 236-244.
C
C  THE PARTICULAR ALGORITHM USED HERE IS DESCRIBED IN
C    WISHART, D., AN ALGORITHM FOR HIERARCHICAL CLASSIFICATIONS,
C    BIOMETRICS, VOLUME 22, NUMBER 1, MARCH 1969, PP 165-170.
C
       DIMENSION S(1),NEAR(1),SREF(1),LIST(1),NUMBR(1),SUM(1)
       IF ( JOB - 2 ) 10, 25, 30
C  JOB=1, INITIALIZE.
C  NUMBR(I)=NUMBER OF ENTITIES CURRENTLY IN THE I-TH CLUSTER
10     WRITE(6,2000)
2000   FORMAT(44HOHIERARCHICAL GROUPING BY THE METHOD OF WARD)
       DO 20 J=1,N
20     NUMBR(J)=1
       W=0.
       BIG=SIGN*1.E50
       RETURN
```

```
C  JOB=2, CALCULATE OBJECTIVE FUNCTION VALUE
25    W=W+SREFX/2.
      SREFX=W
      RETURN
C  JOB=3, UPDATE FOR NEXT ROUND.
30    LRET=LFIND(LREF,NREF)
      NTOT=NUMBR(LREF)+NUMBR(NREF)
      DO 40 J=1,NCL
      I=LIST(J)
      IF(I.EQ.NREF) GO TO 40
C  RECALL THAT LREF HAS BEEN REMOVED FROM LIST SO I NEED NOT BE
C  TESTED FOR EQUALITY WITH LREF.
      LL=LFIND(I,LREF)
      LN=LFIND(I,NREF)
      S(LN)=(S(LN)*(NUMBR(I)+NUMBR(NREF))+S(LL)*(NUMBR(I)+NUMBR(LREF))-
     AS(LRET)*NUMBR(I))/(NTOT+NUMBR(I))
40    CONTINUE
      NUMBR(NREF)=NTOT
C  UPDATE THE NEAR AND SREF ARRAYS.  IF THE EXTREME ELEMENT IN ROW I
C  WAS EITHER LREF OR NREF, THEN IT IS NECESSARY TO FIND A NEW EXTREME
C  ELEMENT.  ROWS PRIOR TO NREF NEED NOT BE CONSIDERED.
      DO 50 J=1,NCL
      I=LIST(J)
      IF(I.EQ.NREF) GO TO 55
50    CONTINUE
55    IF(J.EQ.1) GO TO 80
60    SREF(I)=BIG
      J1=J-1
      DO 70 L=1,J1
      LISTL=LIST(L)
      LL=LFIND(I,LISTL)
      IF(((S(LL)-SREF(I))*SIGN).GE.0.) GO TO 70
      NEAR(I)=LISTL
      SREF(I)=S(LL)
70    CONTINUE
80    J=J+1
      IF(J.GT.NCL) RETURN
      I=LIST(J)
      IF(NEAR(I).EQ.LREF.OR.NEAR(I).EQ.NREF) GO TO 60
      GO TO 80
      END
*EOR
      PROGRAM CLUSVEC(INPUT = 2038, OUTPUT = 2038, TAPE1, TAPE8,
     1 TAPE5 = INPUT, TAPE6 = OUTPUT)
C
C     THIS PROGRAM PRODUCES A COMPLETE PAIRWISE DISTANCE VECTOR FROM THE
C     STEP BY STEP MERGE INFORMATION PROVIDED BY THE CLUSTER PROGRAMS.
C     THE DISTANCES OUTPUT ARE THE FINALIZED DISTANCES BETWEEN CASES AS
C     DETERMINED BY THE CLUSTERING ALGORITHM.
C        FILE 1 IS THE INPUT FILE CONTAINING THE MERGE INFORMATION AS
C     PROVIDED BY THE CLUSTERING PROGRAM.
C        THE CORRESPONDENCE BETWEEN THE NAMES USED FOR THE INPUT VARI-
C     ABLES IN THIS PROGRAM AND THEIR NAMES IN THE CLUSTER PROGRAM OUT-
C     PUT STATEMENTS IS AS FOLLOWS...
C          IDAT(1, K).........K
C          IDAT(2, K).........I(K)
C          IDAT(3, K).........J(K)
C          IDAT(4, K).........S(K)
C          IDAT(5, K).........IL(K)
C          IDAT(6, K).........JL(K)
C           NOTE...IDAT(4, K) CONTAINS A REAL QUANTITY.  IT IS, HOWEVER,
C           ONLY READ AND WRITTEN, AND NOT USED IN ANY COMPUTATIONS.
C        FILE 8 IS THE OUTPUT FILE WHERE THE PAIRWISE DISTANCE
C     VECTOR IS WRITTEN.  THIS FILE CONTAINS ONE LINE PER PAIR, EACH
C     LINE CONTAINING THE SEQUENCE NUMBERS OF THE CASES IN THE PAIR,
C     AND THE DISTANCE FOR THAT PAIR, UNDER THE FORMAT (2I5,F12.7).
C     NOTE THAT THIS FILE IS NOT SORTED AS WRITTEN BY PROGRAM CLUSVEC.
C     THE FILE IS SORTED BY A SYSTEM SORT ROUTINE, CALLED SORT/MERGE
C     AT THE UNIVERSITY OF WASHINGTON CDC 6000 INSTALLATION.  THE FILE
C     IS SORTED SO THAT THE ORDER OF THE VECTOR IS IN THE ORDER OF A
C     SEQUENTIALLY STORED LOWER LEFTHAND TRIANGULAR MATRIX, I.E. THE
C     PAIRS ARE IN ASCENDING ORDER AND THE FIRST ELEMENT OF THE PAIR IS
C     OF HIGHER SIGNIFICANCE THAN THE SECOND ELEMENT OF THE PAIR.
C        THIS PROGRAM USES DYNAMIC STORAGE ALLOCATION, AND READS THE
```

```
C      NUMBER OF CASES FROM THE INPUT (OR RCPAR) FILE.
C      PROGRAM WAS WRITTEN BY R.L. FLEWELLING, UNIV. OF WASH., AND MODI-
C      FIED BY ROBERT B. LEDINGHAM, UNIV. OF WASH.
C
       COMMON IDAT(8, 1)
       DATA LMEM / 4LMEMP /
       READ (5, 200) NE
  200  FORMAT(4X,I4)
       MEMORY = ISHIFT(LOCF(IDAT(1, 1)) + 8 * NE, 30)
       CALL RAP1(LMEM + LOCF(MEMORY))
       NC=NE-1
       DO 5 I=1,NC
       READ (1, 10) (IDAT(J, I), J = 1, 6)
10     FORMAT(3I5,E16.8,2I5)
5      CONTINUE
20     CX=NC
21     DO 23 I = 1, NE
       IDAT(7, I) = 0
23     IDAT(8, I) = 0
       IGRP=1
       NOWI=1
       IDAT(7, IGRP)=IDAT(2, CX)
       IF(IDAT(5, CX).EQ.0) GO TO 50
       NCM=CX-1
28     DO 30 NLM=1,NCM
       NL=NCM-NLM+1
       IF(IDAT(2, NL).NE.IDAT(7, NOWI)) GO TO 30
       IGRP=IGRP+1
       IDAT(7, IGRP)=IDAT(3, NL)
30     CONTINUE
       NOWI=NOWI+1
       IF(IDAT(7, NOWI).EQ.0) GO TO 50
       GO TO 28
50     JGRP=1
       NOWJ=1
       IDAT(8, JGRP)=IDAT(3, CX)
       IF(IDAT(6, CX).EQ.0) GO TO 100
       NCM=CX-1
55     DO 60 NLM=1,NCM
       NL=NCM-NLM+1
       IF(IDAT(2, NL).NE.IDAT(3, NOWJ)) GO TO 60
       JGRP=JGRP+1
       IDAT(8, JGRP)=IDAT(3, NL)
60     CONTINUE
       NOWJ=NOWJ+1
       IF(IDAT(8, NOWJ).EQ.0) GO TO 100
       GO TO 55
100    DO 90 I=1,NE
       IF(IDAT(7, I).EQ.0) GO TO 95
       DO 85 J=1,NE
       IF(IDAT(8, J).EQ.0) GO TO 90
       M1=IDAT(7, I)
       M2=IDAT(8, J)
       MN=MINO(M1,M2)
       MX=MAXO(M1,M2)
       WRITE(9,91)MX,MN,IDAT(4, CX)
91     FORMAT(2I5,F12.7)
85     CONTINUE
90     CONTINUE
95     CX = CX - 1
       IF(CX.GT.0)GO TO 21
       STOP
       END
*EOR
       PROGRAM CORSTD(INPUT = 2038, TAPE2 = 10038, TAPE1 = 10038, TAPE8,
      1 TAPE11, TAPE12, TAPE13, TAPE14, TAPE15, TAPE16, TAPE17 = 10038,
      2 TAPE47, OUTPUT = 2038, TAPE5 = INPUT, TAPE6 = OUTPUT)
       DIMENSION MNAM(7), ODJL(8), ODIN(10), DJS(8), DJSS(8), IX(1), T(8)
      1 , SCORR(8, 8), W(7), AKX(1), DJM(10)
       COMMON NS, IS(2, 12), KX(2047), KXM, X(1)
       EQUIVALENCE (IX(1), X(1)), (AKX(1), KX(1))
       DATA DJS, DJSS / 16 * 0.0 /
       DATA LMEAN, LSD, LM000, LMEMP, IBL / 4HMEAN, 7HSTD DEV, 4HM000,
      1 4LMEMP, 1H  /
```

```
      INT5P(Z) = MINO(MAXO(IFIX(SIGN(ABS(Z) * 1.E5 + 0.5, Z)),
     1 -9900000), 9900000)
C
C     FROM FILES 11..NMETH + 10 GET PAIRWISE JOINING DISTANCES FOR THE
C     NMETH METHODS.  FROM FILE 1 GET ORIGINAL DISTANCE MATRIX.  DO
C     SPEARMAN CORRELATIONS AND WRITE COMPOSITE DISTANCE MATRIX TO
C     FILE 2, WHERE COMPOSITE DISTANCE IS WEIGHTED SUM OF JOINING
C     DISTANCES AFTER STANDARDIZATION, WITH THE WEIGHTS THE SQUARED
C     CORRELATIONS BETWEEN ORIGINAL DISTANCE AND JOINING DISTANCES.
C     IF SPECIFIED, WRITE CORRELATION MATRIX AND STANDARDIZED DISTANCES
C     TO FILE 3.  FILE 47 IS SCRATCH FILE.  NMETH SHOULD BE < 8, AND
C     CORRELATION MAY BE PERFORMED ON SAMPLE OF < 2048 SETS OF DISTANCES
C     TO INCREASE SAMPLE SIZE BEYOND 2047, SEE INSTRUCTIONS IN S.R. SORT
C     THIS PROGRAM WAS WRITTEN BY ROBERT B. LEDINGHAM, UNIV. OF WASH.
C
C     READ PARAMETERS, GET SPACE FOR DATA FOR CORRELATION.
C     FOR SAMPLING PARAMETER SAM...IF SAM .LE. 0.0, SET SAM TO USE ABOUT
C     600 PAIRS.  ADJUST SAM DOWNWARD IF NOT ENOUGH MEMORY AVAILABLE.
C     START OFF BY ALLOCATING ENOUGH MEMORY FOR ABOUT 1000 VALUES.
C
      CALL SECOND(T0)
      READ (5, 200) NH, NMETH, SAM, ISJL, MNAM
      ISJL = MINO(IABS(ISJL), 1)
      MM = 2 - ISJL
      IF ( NMETH .LT. 1 ) NMETH = 6
      NMP1 = NMETH + 1
      DO 20 M = 1, NMETH
   20 IF ( MNAM(M) .EQ. IBL ) MNAM(M) = LM000 + ISHIFT(M, 36)
      WRITE(6, 210) NH, SAM, ISJL, NMETH, MNAM, T0
      FNN = NN = (NH - 1) * NH / 2
      FNNM = FNN - 1.0
      IF ( SAM .LE. 0.0 ) SAM = 600.0 / FNN
      SAM = AMIN1(SAM, 1.0)
      FNVAL = FLOAT(NMP1) * FNN
C
C     ...GET JOBCARD MAXIMUM MEMORY ALLOWABLE IN MAVAIL.  SUBROUTINE
C     ...RAP1 PERFORMS SYSTEM CALL TO PP ROUTINE MEM.
C
      MAVAIL = ISHIFT(-1, 30) .AND. MASK(30)
      CALL RAP1(LMEMP + LOCF(MAVAIL))
      MAVAIL = ISHIFT(MAVAIL, -30) - LOCF(X(1))
      IF ( MAVAIL .GE. IFIX(FNVAL * SAM * 1.1) ) GO TO 30
      SAM = FLOAT(MAVAIL) / (FNVAL * 1.1)
      WRITE(6, 211) MAVAIL
   30 ISAMP = SAM + 0.01
      MREQ = ((MINO(IFIX(FNVAL * SAM * 0.9), 1063) + LOCF(X(1)))/64)*64
C
C     ...MREQ IS TOTAL REQUIRED FIELD LENGTH (FOR NOW), NXA IS AVAILABLE
C     ...MEMORY FOR X ARRAY (DATA FOR CORRELATIONS).  USE RAP1 AGAIN TO
C     ...GET MEMORY.  INITIALIZE RANDOM NUMBER GENERATOR.
C
      NXA = MREQ - LOCF(X(1)) - NMP1
      MEMORY = ISHIFT(MREQ, 30)
      CALL RAP1(LMEMP + LOCF(MEMORY))
      CALL RANSET(T0)
      NXC = NS = 0
      I = 11
C
C     READ DATA, SAVE ON FILE 47, ALSO SAVE SAMPLE IN MEMORY, GETTING
C     MORE MEMORY IF NEEDED IN 512-WORD CHUNKS.  ACCUMULATE SUM AND SUM
C     OF SQUARES FOR EACH DISTANCE FOR ALL CASES.
C
      DO 75 J = 1, NN
      I = I + 1
      IF ( I .LT. 11 ) GO TO 40
      READ (1, 250) ODIN
      I = 1
   40 ODJL(1) = ODIN(I)
      DO 50 M = 1, NMETH
      K = M + 10
   50 READ (K, 230) ODJL(M+1)
      WRITE(47) (ODJL(M), M = MM, NMP1)
      IF ( ISAMP .NE. 0 ) GO TO 55
      IF ( RANF(0.0) .GE. SAM ) GO TO 70
```

```
   55 IF ( NS .GE. 2047 ) GO TO 57
      IF ( NXC .LT. NXA ) GO TO 60
      NXA = NXA + 512
      MREQ = MREQ + 512
      IF ( NXA .LE. MAVAIL ) GO TO 58
   57 WRITE(6, 213) NS, J, NN, MREQ
      SAM = -1.0
      ISAMP = 0
      GO TO 70
   58 MEMORY = ISHIFT(MREQ, 30)
      CALL RAP1(LMEMP + LOCF(MEMORY))
   60 NS = NS + 1
      DO 65 M = 1, NMP1
      NXC = NXC + 1
   65 X(NXC) = ODJL(M)
   70 DO 75 M = 1, NMP1
      OJS(M) = ODJL(M) + OJS(M)
   75 OJSS(M) = ODJL(M) ** 2 + OJSS(M)
C
C     HAVE DATA FOR CORRELATIONS IN X(1..NMP1, 1..NS).  SPEARMAN CORREL-
C     ATION PROCEDURE USES FORMULAS FROM SPSS MANUAL, 2ND EDITION,
C     PP 239-290.
C     START CORRELATIONS BY, FOR EACH SET OF DISTANCES, REPLACING VALUES
C     BY ORDINAL RANKS, KEEPING TRACK OF CORRECTION FOR TIES (T).  IN
C     THE CASE OF TIES, ASSIGN AVERAGE RANK TO EACH TIED ELEMENT.
C     RECALL THAT X AND IX ARRAYS ARE EQUIVALENCED, AND KX AND AKX
C     ARRAYS ARE EQUIVALENCED.
C
      NNNS = (NS ** 2 - 1) * NS
      NSM = NS * NMP1
      PRINT 212, SAM, NS, MREQ
      MO = -NMP1
      DO 100 M = 1, NMP1
      MO = MO + 1
      K = M
      ITE = NNNS
C
C     ...MAKE ARRAY KX...KX(I = 1..NS) = IX(M, I = 1..NS) WITH LOW-ORDER
C     ...12 BITS OF KX(I) REPLACED BY VALUE I.  THEN SORT KX ARRAY.
C
      DO 85 I = 1, NS
      KX(I) = IX(K) .AND. MASK(48) .OR. I
   85 K = K + NMP1
      CALL SORT
C
C     ...REPLACE X(M, 1..NS) WITH ORDINAL RANKS.  ITI IS NUMBER OF
C     ...INSTANCES OF VALUE RCO.  LOOKING AT EACH ELEMENT OF KX(SORTED
C     ...ARRAY) IN TURN...
C
      ITI = 0
      RCO = -1.E99
      DO 95 I = 1, NS
      XI = AKX(I) .AND. MASK(48)
      IF ( RCO .LT. XI ) GO TO 90
C
C     ...VALUE SAME AS PREVIOUS.  INCREMENT ITI.
C
      ITI = ITI + 1
      GO TO 95
C
C     ...VALUE HAS CHANGED.  FOR THE ITI PREVIOUS ELEMENTS, PUT IN
C     ...(AVERAGE) RANK XR, AND INCREMENT T ACCUMULATOR IF ITI > 1.
C
   90 RCO = XI
      IF ( ITI - 1 ) 94, 91, 92
   91 K = (KX(I-1) .AND. 77778) * NMP1 + MO
      X(K) = I - 1
      GO TO 95
   92 XR = FLOAT(2 * I - ITI - 1) * 0.5
      DO 93 J = 1, ITI
      K = (KX(I-J) .AND. 77778) * NMP1 + MO
   93 X(K) = XR
      ITE = ITE - (ITI ** 2 - 1) * ITI
   94 ITI = 1
```

```
   95 CONTINUE
C
C     ...REPLACE FINAL ITI ELEMENTS WITH RANK, AS ABOVE, AND COMPUTE T
C
      XR = FLOAT(2 * NS - ITI + 1) * 0.5
      DO 97 J = 1, ITI
      K = (KX(NS - J + 1) .AND. 7777B) * NMP1 + M0
   97 X(K) = XR
      IF ( ITI .GE. 2 ) ITF = ITF - (ITI ** 2 - 1) * ITI
      T(M) = FLOAT(ITF) / 12.0
  100 CONTINUE
C
C     COMPUTE SPEARMAN CORRELATIONS FOR EACH PAIR...PRINT CORRELATIONS,
C     AND WRITE THEM TO FILE 9 IF ISJL .NE. 0 .
C
      DO 120 M1 = 1, NMETH
      M2M = M1 + 1
      SCORR(M1, M1) = 1.0
      DO 120 M2 = M2M, NMP1
      K1 = M1
      RD = 0.0
      DO 110 K2 = M2, NSM, NMP1
      RD = (X(K1) - X(K2)) ** 2 + RD
  110 K1 = K1 + NMP1
  120 SCORR(M1, M2) = SCORR(M2, M1) = (T(M1) + T(M2) - RD) * 0.5
     1 / SQRT(T(M1) * T(M2))
      SCORR(NMP1, NMP1) = 1.0
      WRITE(6, 215) MNAM, (SCORR(K, 1), K = 1, NMP1)
      DO 130 M = 2, NMP1
  130 WRITE(6, 216) MNAM(M-1), (SCORR(K, M), K = 1, NMP1)
      IF ( ISJL .EQ. 0 ) GO TO 140
      DO 135 M = 1, NMP1
  135 WRITE(9, 280) (SCORR(K, M), K = 1, NMP1)
C
C     GET MEAN, STD DEV FOR DISTANCES.  FOR EACH CASE, STANDARDIZE VARI-
C     ABLES, WRITE THEM TO FILE 8 IF ISJL .NE. 0, AND WRITE COMPOSITE
C     DISTANCE SCORE IN FORMAT (10F10.6), I.E. 10 CASES PER LINE.
C     FILE 2 IS OUTPUT FILE FOR COMPOSITE DISTANCE VECTOR.
C
  140 DO 145 M = 1, NMETH
  145 W(M) = SCORR(M+1, 1) ** 2
      WRITE(6, 217) MNAM, (W(M), M = 1, NMETH)
      DO 150 M = 1, NMP1
      OJM(M) = OJS(M) / FNN
  150 OJSS(M) = SQRT((OJSS(M) - OJS(M) * OJM(M)) / FNNM)
      WRITE(6, 218) LMEAN, (OJM(M), M = 1, NMP1)
      WRITE(6, 218) LSD, (OJSS(M), M = 1, NMP1)
      REWIND 47
      I = 0
      IF ( ISJL .NE. 0 ) GO TO 170
C
C     ...ISJL = 0... DON'T WRITE Z-SCORES.  BINARY FILE 47 HAS DISTANCES
C     ...FOR THE NMETH METHODS ONLY.
C
      DO 165 K = 1, NN
      READ (47) (ODJL(M), M = 1, NMETH)
      WJL = (ODJL(1) - OJM(2)) * W(1) / OJSS(2)
      DO 160 M = 2, NMETH
  160 WJL = (ODJL(M) - OJM(M+1)) * W(M) / OJSS(M+1) + WJL
      I = I + 1
      ODIN(I) = WJL
      IF ( I .LT. 10 ) GO TO 165
      I = 0
      WRITE(2, 250) ODIN
  165 CONTINUE
      GO TO 190
C
C     ...ISJL .NE. 0... WRITE Z-SCORES TO FILE 8. BINARY FILE 47 HAS
C     ...ORIGINAL DISTANCES AND JOINING DISTANCES FOR THE NMETH METHODS.
C
  170 DO 180 K = 1, NN
      READ (47) (ODJL(M), M = 1, NMP1)
      KX(1) = INTSP((ODJL(1) - OJM(1)) / OJSS(1))
      OJ = (ODJL(2) - OJM(2)) / OJSS(2)
```

```
      WJL = DJ * W(1)
      KX(2) = INT5P(DJ)
      DO 175 M = 3, NMP1
      DJ = (ODJL(M) - DJM(M)) / DJSS(M)
      WJL = DJ * W(M-1) + WJL
  175 KX(M) = INT5P(DJ)
      I = I + 1
      ODIN(I) = WJL
      IF ( I .LT. 10 ) GO TO 190
      I = 0
      WRITE(2, 250) ODIN
  180 WRITE(8, 282) K, (KX(M), M = 1, NMP1)
C
C     ...WRITE LAST LINE OF COMPOSITE SCORES TO FILE 2, IF
C     ...NH * (NH - 1) / 2 MOD 10 .NE. 0, AND PRINT ELAPSED TIMES.
C
  190 IF ( I .NE. 0 ) WRITE(2, 250) (ODIN(K), K = 1, I)
      CALL SECOND(T1)
      T10 = T1 - T0
      WRITE(6, 219) T1, T10
      STOP
C
  200 FORMAT(4X,I4,12X,I4,F4.2,4X,I4,4X,7A4)
  210 FORMAT(21H1NUMBER OF ENTITIES =I4/35H SAMPLING PARAMETER (INPUT VA
     1LUE) =F5.2/46H STANDARDIZED JOINING DISTANCE OUTPUT OPTION =I4,22H
     2   NUMBER OF METHODS =I4/20H NAMES OF METHODS...7A6/24H ELAPSED TI
     3ME AT START =F10.3,8H SECONDS)
  211 FORMAT(50HOSAMPLING PARAMETER ADJUSTED TO FIT JOB CARD CM = D6)
  212 FORMAT(26HOSAMPLING PARAMETER USED =F5.2,7H GIVINGI6,17H PAIRS REQ
     1UIRING D6,16H WORDS OF MEMORY)
  213 FORMAT(36H-*** WARNING...MEMORY OVERFLOW AFTER16,15H CASES SAVED..
     1./20H *** CASES SAMPLED =I8,3H OFI8,19H...MEM REQUESTED = D6)
  215 FORMAT(38H-COPHENETIC (SPEARMAN) CORRELATIONS.../1H08X8HD.D.
     1 7A8/5HOD.D.8F8.4)
  216 FORMAT(1X,A4,8F8.4)
  217 FORMAT(1H-11X,8HD.D.     7A8/7HOWEIGHT9X,7F8.4)
  218 FORMAT(1HOA7,8F9.4)
  219 FORMAT(25H-ELAPSED TIME AT FINISH =F11.3,8H SECONDS/26H TIME FOR T
     1HIS PROCEDURE =F10.3,8H SECONDS)
  230 FORMAT(10X,F12.7)
  250 FORMAT(10F10.6)
  280 FORMAT(8F10.7)
  282 FORMAT(I6,10X,8I8)
      END
*FOR
      SUBROUTINE SORT
      COMMON NS, IS(2, 12), KX(2047), KXM
      REAL KX, KXM
      INTEGER F
C
C     SORT ARRAY KX(1..NS) INTO ASCENDING ORDER USING QUICKSORT, ALGOR-
C     ITHM 7.4 OF REINGOLD, NIEVERGELT, AND DEO, MODIFIED TO DO ORDINARY
C     BUBBLE SORT ON SUBARRAYS OF < 8 ELEMENTS.
C     IS(2, 12) IS STACK, WITH POINTER ISP.  MAXIMUM ARRAY SIZE IS
C     2047, WHICH CAN BE INCREASED BY INCREASING DIMENSION OF KX
C     (TO MAX SIZE DESIRED) AND IS (TO LOG2 OF MAX SIZE) IN THIS SUBROU-
C     TINE AND IN MAIN PROGRAM.
C
      IS(1, 1) = IS(2, 1) = 0
      ISP = 1
      KX(NS+1) = 1.E99
      F = 1
      L = NS
C
C     WHILE F < L DO...IF L - F < 8, DO BUBBLESORT, THEN POP THE STACK
C
   10 IF ( L - F .GE. 8 ) GO TO 30
      LM1 = L - 1
      DO 20 I = F, LM1
      DO 20 J = I, LM1
      IF ( KX(J+1) .GE. KX(I) ) GO TO 20
      XT = KX(I)
      KX(I) = KX(J+1)
      KX(J+1) = XT
```

```
   20 CONTINUE
      F = IS(1, ISP)
      L = IS(2, ISP)
      ISP = ISP - 1
      IF ( F .LT. L ) GO TO 10
      RETURN
C
C     ELSE PARTITION THE ARRAY
C
   30 I = RANF(0.0) * FLOAT(L - F) + F
      XF = KX(I)
      KX(I) = KX(F)
      KX(F) = XF
      I = F
   35 I = I + 1
      IF ( KX(I) .LT. XF ) GO TO 35
      J = L
      IF ( KX(J) .LE. XF ) GO TO 45
   40 J = J - 1
      IF ( KX(J) .GT. XF ) GO TO 40
   45 IF ( I .GE. J ) GO TO 65
   50 XT = KX(I)
      KX(I) = KX(J)
      KX(J) = XT
   55 I = I + 1
      IF ( KX(I) .LT. XF ) GO TO 55
   60 J = J - 1
      IF ( KX(J) .GT. XF ) GO TO 60
      IF ( I .LT. J ) GO TO 50
   65 KX(F) = KX(J)
      KX(J) = XF
C
C     TAKE APPROPRIATE ACTION, DEPENDING ON WHICH, IF ANY, SUBARRAYS ARE
C     NONTRIVIAL
C
      IF ( F .LT. J - 1 ) GO TO 75
      IF ( J + 1 .LT. L ) GO TO 70
C
C     BOTH ARE TRIVIAL...POP THE STACK
C
      F = IS(1, ISP)
      L = IS(2, ISP)
      ISP = ISP - 1
      IF ( F .LT. L ) GO TO 10
      RETURN
C
C     RIGHT SUBARRAY ONLY IS NON-TRIVIAL...SORT IT
   70 F = J + 1
      GO TO 10
C
   75 IF ( J + 1 .LT. L ) GO TO 80
C
C     LEFT SUBARRAY ONLY IS NON-TRIVIAL...SORT IT
C
      L = J - 1
      GO TO 10
C
C     BOTH SUBARRAYS ARE NON-TRIVIAL...PUT LARGER ONE ON STACK, SORT
C     OTHER ONE
C
   80 IF ( L - J .LT. J - F ) GO TO 85
      ISP = ISP + 1
      IS(1, ISP) = J + 1
      IS(2, ISP) = L
      L = J - 1
      GO TO 10
   85 ISP = ISP + 1
      IS(1, ISP) = F
      IS(2, ISP) = J - 1
      F = J + 1
      GO TO 10
      END
*EOR
*EOF
```

```
QBOBL,T900,IO600,CM120K,P1.
ACCOUNT.
ATTACH,RCBINF,RCBF2,ID=RCRF2.MR=1.
COPYR,RCBINF,METH.
ATTACH,FSIN,FS3194.
ATTACH,TAPE9,L3L194,MR=1.
MAP,OFF.
METH,,FSIN.
CATALOG,RCPAR,RP3194,RP=0.
RETURN,FSIN.
COPYR,RCBINF,RCMAIN.6.
SKIPF,RCBINF,1.
COPYR,RCBINF,RCMETH.6.
COPYR,RCBINF,CLUSVEC.
REWIND,RCMETH.
DUP.6,*.
REWIND,TAPE2,TAPE7,TAPE9,RCPAR,METH.
COPYR,RCMETH,METH.
LOAD,RCMAIN.
METH,RCPAR,DUMOUT.
REWIND,TAPE7,TAPE8,RCPAR.
CLUSVEC,RCPAR,,TAPE7.
FILE,TAPE8,BT=C,RT=Z,FL=136.
FILE,TAPE1$,BT=C,RT=Z,FL=136.
REWIND,TAPE8,TAPE1$.
SORTMRG,O=DUMOUT.
CATALOG,TAPE1$,JD3194$,RP=0.
ENDUP.
COPYR,RCBINF,LGO,?.
REWIND,TAPE2,TAPE8,TAPE11,TAPE12,TAPE13.
REWIND,TAPE14,TAPE15,TAPE16,RCPAR.
LGO,RCPAR,CJM,TAPE2.
CATALOG,CJM,VC3194,RP=0.
REWIND,RCMAIN,RCMETH,TAPE9,METH,CJM,RCPAR.
COPYR,RCMETH,METH.
LOAD,RCMAIN.
METH,RCPAR,COMPOUT,CJM,MERGOUT.
REWIND,COMPOUT,MERGOUT.
COPYE,COMPOUT,OUTPUT.
COPYSF,MERGOUT,OUTPUT.
CATALOG,COMPOUT,OUT3194,RP=0.
DUP,6,1.
PURGE,TAPE1$.
PURGE,RCPAR.
EXIT.
```

```
REWIND,DUMOUT.
COPYE,DUMOUT,OUTPUT.
*EOR
   1 194   6   1   0   5 .10   0   0   ICL  AR  AW  MED CEN WARD
.267       .251       .039       .446       .066       .145
(3F10.6,10X3F10.6)
194 HOSPITALS, ENDOGENOUS REGRESSION, 5 ESCORES + TARG OCC
*EOR
SORT
FILE,INPUT=TAPE8,OUTPUT=TAPE11
FIELD,KEY1(1,5,DISPLAY),KEY2(6,5,DISPLAY)
KEY,KEY1,KEY2
END
*EOR
SORT
FILE,INPUT=TAPE8,OUTPUT=TAPE12
FIELD,KEY1(1,5,DISPLAY),KEY2(6,5,DISPLAY)
KEY,KEY1,KEY2
END
*EOR
SORT
FILE,INPUT=TAPE8,OUTPUT=TAPE13
FIELD,KEY1(1,5,DISPLAY),KEY2(6,5,DISPLAY)
KEY,KEY1,KEY2
END
*EOR
SORT
FILE,INPUT=TAPE8,OUTPUT=TAPE14
FIELD,KEY1(1,5,DISPLAY),KEY2(6,5,DISPLAY)
KEY,KEY1,KEY2
END
*EOR
SORT
FILE,INPUT=TAPE8,OUTPUT=TAPE15
FIELD,KEY1(1,5,DISPLAY),KEY2(6,5,DISPLAY)
KEY,KEY1,KEY2
END
*EOR
SORT
FILE,INPUT=TAPE8,OUTPUT=TAPE16
FIELD,KEY1(1,5,DISPLAY),KEY2(6,5,DISPLAY)
KEY,KEY1,KEY2
END
*EOR
*EOF
```

REFERENCES

Anderberg, Michael R. Cluster Analysis for Applications. Academic Press, New York (Spring, 1973).

Ball, Geoffrey H., and David Hall. "ISODATA, A Novel Method of Data Analysis and Pattern Classification," Technical Report, Information Sciences Branch, Office of Naval Research, April, 1965.

Bauer, Katherine. "Classifying Hospitals for Purposes of Prospective Reimbursement," Harvard Center for Community Health and Medical Care, August, 1974.

Berry, Ralph. "Cost and Efficiency in the Production of Hospital Services," Health and Society, 52:291-313, Summer, 1974.

Berry, Ralph. "On Grouping Hospitals for Economic Analysis," Inquiry, 10:5-12, December, 1973.

Boyce, A. J. "Mapping Diversity: A Comparative Study of Some Numerical Methods," Numerical Taxonomy (A. J. Cole, ed.), Academic Press, London, 1969.

Clarkson, Kenneth. "Property Rights: Some Implications for the Non-Profit Hospital," Journal of Law and Economics, 15:363-384, October, 1972.

Cormack, R. M. "A Review of Classification," Journal of the Royal Statistical Society (Series A), 134:321-353, 1971.

Daling, Janet R., and H. Tamura. "Use of Orthogonal Factors for Selection of Variables in a Regression Equation -- An Illustration," The Journal of the Royal Statistical Society, Series C (Applied Statistics), 19:260-268, 1970.

Dowling, William L., et al. "Evaluation on Prospective Reimbursement of Hospitals in Downstate New York," Center for Health Services Research, University of Washington, December, 1976.

Dowling, William L. "Prospective Reimbursement of Hospitals," Inquiry, Vol. 11, No. 3, September, 1974.

Doyle, Peter, and Ian Fenwick. "The Pitfalls of AID Analysis," The Journal of Marketing Research, 12:408-413, November, 1975.

Edwards, Miller, and Schumacher. "Classification of Community Hospitals by Scope of Services: Four Indices," Health Services Research, Winter, 1972.

Feldstein, Martin. "Econometric Studies in Health Economics," Frontiers of Quantitative Economics (Michael Intrilligator and David Kendrick, eds.), North Holland Press, Amsterdam, 1974.

Feldstein, Martin. The Rising Cost of Hospital Care. Information Resources Press, Washington, 1971.

Feller William. An Introduction to Probability Theory and Its Applications, Volume I. John Wiley and Sons, Inc., New York, NY, 1968.

Fisher, Walter D. "On Grouping for Maximum Homogeneity," Journal of the American Statistical Association, 53:789-798, December, 1958.

Garfinkle, Robert, and G. L. Nemhauser. Integer Programming. John Wiley and Sons, Inc., New York, NY, 1972.

Gower, J. C. "A Comparison of Some Methods of Cluster Analysis," Biometrics, 624-637, December, 1967.

Green, Paul E., and Frand J. Carmone. Multidimensional Scaling and Related Techniques in Marketing Analysis. Allyn and Bacon, Inc., Boston, MA, 1970.

Harmon, H. Modern Factor Analysis. University of Chicago Press, Chicago, IL, 1968.

Hawkins, Douglas M. "On the Investigation of Alternative Regressions by Principal Components Analysis," The Journal of the Royal Statistical Society, Series C (Applied Statistics), 20:275-286, 1972.

Klastorin, T. D. "A Clustering Approach to Systems Design," Unpublished Ph.D. Dissertation, The University of Texas at Austin, Austin, TX, 1973.

Lance, G. N., and W. T. Williams. "Computer Programs for Hierarchical Polythetic Classification (Similarity Analyses)," Computer Journal, 9:60-64, May, 1966.

Lave, Judith, and Lester Lave. "Hospital Cost Functions," American Economic Review, 60:379-395, June, 1970.

Lee, Mah Lin. "A Conspicuous Production Theory of Hospital Behavior," Southern Economic Journal, 38:48-58, July, 1971.

MacQueen, James B. "Some Methods for Classification and Analysis of Multivariate Observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, June 21 - July 18, 1965, and December 27 - January 7, 1966; University of California Press, 1967.

Mahalanobis, P. C. "On the Generalized Distance in Statistics," Proceedings of the National Institute of Science in India, 2:49-58, 1936.

Morgan, James N., and John A. Sonquist. "Problems in the Analysis of Survey Data and a Proposal," Journal of the American Statistical Association, 58:415-434, September, 1963.

Newhouse, Joseph. "Toward a Theory of Non-Profit Institutions: An Economic Model of a Hospital," American Economic Review, 60:60-74, 1970.

Noll, Roger. "The Consequences of Public Utility Regulation of Hospitals," Unpublished working paper, California Institute of Technology, 1973.

Peterson, Robert A. "Sequential Cluster Analysis in Market Structuring," Working Paper #72-18, Graduate School of Business, The University of Texas at Austin, Austin, TX, December, 1971.

Phillip, P. Joseph, and Ramani N. Iyer. "Classification of Community Hospitals," Health Services Research, 349-350, Winter, 1972.

Phillip, P. Joseph, and Ramani N. Iyer. "A Taxonomy of Community Hospitals," Unpublished report, The Bureau of Research Statistics, The American Hospital Association, Chicago, IL, June, 1974.

Pointer, D., and J. Phillip. "Classifying Short-Term Hospitals for Routine Service Cost Limitation Under Section 223 of P. L. 92-603: A Critical Analysis," Unpublished paper, Department of Teaching Hospitals, Association of American Medical Colleges, Washington, D. C., May, 1974.

Rand, William. "Objective Criteria for the Evaluation of Clustering Methods," Journal of the American Statistical Association, 66:846-850, December, 1971.

Rosenthal, Gerald. "The Demand for General Hospital Facilities," Hospital Monograph Series No. 14, American Hospital Association, Chicago.

Salkever, David, and Tom Bice. "The Impact of Certificate of Need Controls on Hospital Investments," The John Hopkins University, August, 1975.

Shortell, Stephen, et al. "The Effects of Management Practices on Hospital Efficiency and Quality of Care," Organizational Research in Hospitals, Inquiry Monograph (Shortell and M. Brown, eds.), Blus Cross Association, September, 1976.

Sneath, P. H. A., and Robert R. Sokal. Numerical Taxonomy. W. H. Freeman and Company, San Francisco, CA, 1973.

Sokal, R. R. "Classification: Purposes, Principles, Progress, Prospects," Science, 185:1115-1119, September, 1974.

Sokal, R. R., and C. D. Michener. "A Statistical Method for Evaluating Systematic Relationships," University of Kansas Scientific Bulletin, 38:1409-1438, 1958.

Sokal, R. R., and F. James Rohlf. "The Comparison of Dendrograms by Objective Methods," Taxon, 11:33-40, February, 1962.

Trivedi, V. M. "Taxonomy of Short-Term General Hospitals for Control and Equity," Inquiry, 1977.

Ward, J. H. "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, 58:236-244, 1963.

Ward, J. H., and M. E. Hook. "Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles," Educational and Psychological Measurement, 23:69-82, 1963.

Wishart, D. "A Generalized Approach to Cluster Analysis," Part of Ph.D. Thesis, University of St. Andrews, 1970.

Worthington, Paul N., and Jesse S. Hixson. "Setting Limits on Reimbursement of Hospital Costs Under Medicare," Unpublished paper, Office of Research and Statistics, Social Security Administration, U. S. Department of Health, Education, and Welfare, Washington, D. C., 1975.

*U.S. GOVERNMENT PRINTING OFFICE: 1979 620-010/4019 1-3